Crystallographic studies of modified insulin

by

Maria Gertrudis Wilhelmina Turkenburg-van Diepen

A Thesis Submitted to the
University of York
for the
Degree of Doctor of Philosophy

Department of Chemistry
University of York
September 1996

Abstract

Eight X-ray crystallographic studies of modified insulins are described, and, where possible, the resulting models are compared with relevant native insulin structures or other modified insulin structures.

Three of the structures, namely B13 Glu→Gln insulin, B9 Ser→His insulin and Zn^{2+} →Co^{2+} insulin, are isomorphous with 4Zn native insulin. The conformation of all three, however, is determined as T3R3^f (the first three residues of the B chains in the R3 trimer extended instead of helical) rather than the T3R3 conformation as described for the native structure. The metal binding of these three modified rhombohedral insulins is different from that of the native structure, which has two axial and three off-axial zinc sites with a total occupancy of 2.67. B13 Glu→Gln insulin lacks the axial zinc in the R3^f trimer. B9 Ser→His insulin contains additional zinc sites, coordinated by the extra histidines in combination with B5 and B10 histidines. Zn^{2+} →Co^{2+} insulin has two axial cobalt ions only, octahedrally coordinated in the T3 trimer and tetrahedrally coordinated in the R3^f trimer. It also contains two phenol molecules in the phenol pocket which is usually seen to contain only one.

Another crystal form of B9 Ser→His insulin, and the structure of B8 Gly→Ser/B13 Glu→Gln/B30 Thr-amide insulin, are isomorphous with monoclinic native insulin. In B9 Ser→His insulin in this crystal form, like in the rhombohedral form of the mutant, additional zinc sites are created with the extra histidines.

The B9 Ser→Asp/B27 Thr→Glu insulin structure is isomorphous with orthorhombic native insulin. Although the B9 Ser→Asp mutation is seen to produce a very stable monomer in solution, the charge repulsion is overcome in the crystal, resulting in dimers in the structure.

The A21 Asn→Gly/B9 Ser→Glu/B10 His→Glu insulin structure is orthorhombic and not isomorphous with any known insulin structure. The conformations of the C-termini of the B chains in the asymmetric dimer have changed to allow for the different, denser, packing.

The structure of A4 Glu→Gln/B25 Phe→Tyr/des B30 single chain insulin is

presumably isomorphous with the structures of rat insulin II and insulin from the snake *Zaocys dhumnades dhumnades Cantor.* Evidence is presented to support the validity of the mutant structure, but the resolution of the data is insufficient to allow detailed analysis.

Acknowledgements

After all those years, I've got quite a list of people who contributed in some way to this thesis, for which I would like to express thanks.

Guy, my supervisor, for the enthusiasm and inspiration, which was always there when I needed it.
Eleanor as a stand-in mother figure, but of course mainly for helping out with all the things a student in our lab needs helping out with.
Johan and Gideon for the supervision that supervisors don't have time for. And for most of the proof-reading. And for being friends.
Jean for being the best of friends, and for forever digging out papers about insulin. And for the many discussions on how to get insulin to behave in our theses.
Garib and Kevin for specific help on the subjects of maximum entropy, maximum likelihood and Lagrange multipliers.
Mike and Kevin for help with LaTeX.
Stuart and Dave E for sheer endless computer problem solving. And Steve and Richard for getting into it, too.
John and Sue for much appreciated last-minute discussions on insulin and John for stepping in with proof-reading. Their timing was amazing!
Zbyszek for much needed last-minute insight on symmetry.
Some people have been invaluable on the practical side of things: Bing, Judy, Mary, Stéphane and Vibeke for crystallisation; Zygmunt, Johan, Gideon, Peter, Jean and Giles for data collection and processing.
Tom for getting X-FIT going at just the right time. It was an absolute pleasure to be one of the "several active crystallographers" (see the X-FIT manual February 1996, p57).
Helen D-S for some very special moments when the going was tough. And Sally and Shirley R, too.

Rag for an enormous amount of faith in me and lots of wise words.
And then there are all the other people who have made York a very special place over all those years: Madeleine, Tony, Keith, Rod, Caroline, Jane, Jinny, Sarah, Anne, Tracey, Alec, Craig, Koen, Freddy, Shirley T, Valérie, Marek, Paul H, Paul E, Leo and Ali, Chandra, Dawn, Camilla, Richard, Lorraine, Rick, Rana, Dave R, Dave L, Teresa, Steve, Leo B, Helen K, Liz, Pavel, Ula, Fritjof, Mike, Will, Julie, Fred, Jeremy, Sara, Colin K and all the others who have passed through.
I would also like to thank my parents for creating an environment in which following this path seemed so natural.
And last but not least Laura, not necessarily for coming along at the right time, but for the very special person she is. And for the incredible amount of patience she had with me in the last six months. It's time to start on that list of things to do "Yes, after your thesis, mummy". In this context Johan needs another special mention, too.

For my children

Contents

List of Figures

List of Tables

Chapter 1

Discovering insulin

1.1 Introduction

The disease Diabetes Mellitus was first described in an Egyptian papyrus, discovered by Ebers in the tomb of Thebes in Egypt in 1862, which is said to have been written between 3000 and 1500 BC. The first use of the term 'Diabetes Mellitus' is accredited to Aretaeus of Cappadocia and Apolonius of Memphis in the second century AD. 'Diabetes' stems from the Greek word for 'pipe-like' because nutrients begin to pass through the system rather than being utilised. 'Mellitus' is Latin for 'honey' or 'sweet', to distinguish the disease from 'Diabetes Insipidus', which is a pituitary disorder in which large volumes of sugar-free urine are passed.

In the middle of the nineteenth century, evidence from autopsies started to suggest a link between the pancreas and Diabetes Mellitus. Diabetics were sometimes seen to have pancreas damage, and patients with damaged pancreases almost always had diabetes. In 1869 Langerhans discovered the existence of two systems of cells in the pancreas: the acinar cells, secreting the pancreatic juice into the digestive system, and islets floating between the acini, with some as yet unknown function. In 1889 Minkowski and Von Mering depancreatised a dog, causing a state of polyuria indistinguishable from diabetes. This was the first direct evidence of

the link between diabetes and the pancreas. They also showed that it was not the absence of the pancreatic juice that caused diabetes by studying the effect of ligating the pancreatic ducts rather than removing the whole pancreas. In most cases this caused minor digestive problems, but never diabetes.

The Frenchman Hédon proved in 1893 that a total pancreatectomy was necessary to cause Diabetes Mellitus. He blocked the flow of pancreatic juice, removed most of the pancreas, and grafted the small pancreatic remains just under the skin of his test subjects, for easy removal at a later stage of the experiment. The blood supply to this piece was left as normal as possible. At this stage no diabetes was established. After removing the graft, Diabetes Mellitus could be diagnosed immediately.

During the 1890s it was discovered that several diseases could be treated by feeding patients extracts of thyroid. In analogy, it was tried extensively, probably by more that 400 researchers worldwide (Bliss, 1982), to treat diabetic patients by feeding them pancreatic extracts, without success. Mildly positive effects could never be reproduced by others. Usually the toxic side-effects were far worse than the positive effects, although sometimes the side-effects were seen as positive, *e.g.* kidney failure could change the urine in such a way that diabetes was no longer diagnosed.

In 1901 Opie showed a direct link between Diabetes Mellitus and damage to the islets of Langerhans, generating a wide belief in an internal secretion in the islets, responsible for the prevention of diabetes. In view of the failed pancreas therapy experiments, scientists suggested that the exocrine pancreatic secretion might destroy the active component of the internal secretion. Two distinct types of pancreas seemed obvious choices for attempts to create pure internal secretion: that from foetuses, in which islets develop before acinar cells, and that from certain types of fish, which have anatomically distinct islet and acinar parts. There are no records to prove the first was tried in the early years, the second was tried between 1902

and 1904, but was not promising. Several scientists all over the world attempted to isolate and purify substances from the pancreas that were supposed to cure Diabetes Mellitus. Meanwhile, several diabetologists kept believing that a diet was the only good way to treat diabetics.

1.2 Earliest treatment of Diabetes Mellitus

Until the 1910s opium was the only widely used medicine in the treatment of Diabetes Mellitus. However, this could only dull the patients' despair, but did nothing to cure or treat. Further treatment consisted of more or less trendy diets. In the late 1850s Piorry advised the use of extra sugar, to compensate for the loss of sugar into the urine. This 'eating a lot to compensate' was practised until the early 1900s. In Paris under German siege, in 1870, Bouchardat noticed that rationing of food caused the disappearance of glycosuria in diabetic patients, while exercise also seemed to have a positive effect. The idea settled that maybe the body should be put under as little metabolic strain as possible by limited eating.

At the time of the earliest tests of pancreatic extracts in the treatment of Diabetes Mellitus, America had two leading diabetologists who did not believe in pancreas therapy. They were Allen and Joslin. They both practised 'starvation treatment' where the patients are undernourished for a certain amount of time. They argued that apart from the carbohydrate metabolism the protein and fat metabolisms in diabetic patients were also affected. By cutting down on food until the patient's body was relieved of all metabolic strain, and then slowly building up again until a reasonably healthy diet was achieved, many diabetics could live years longer. Some patients, however, did not even tolerate the minimum amount of food (the 'living diet'), and succumbed quickly.

1.3 Towards reliable pancreas therapy

As early as the 1890s Paulesco, a Romanian scientist in Paris, developed an interest in the internal secretion of the pancreas. He regained his interest in 1916, when he did his first experiments with extracts in Bucharest. The First World War and Austrian occupation prevented serious experiments until 1919, and he published some successes with his 'pancréine' in 1920 and 1921.

In the early 1900s Zuelzer, in Berlin, developed a pancreatic extract he named 'acomatol' with which he managed to bring back a 50 year old patient from a coma in 1906. This extract, produced for the Schering company, was probably very contaminated and produced many side-effects. It was tested in Minkowski's clinic in 1909, where it was concluded that the positive and negative effects were due to the same component. This caused Zuelzer's funding to be withdrawn, and he stopped publishing. He persisted with his experiments, however, and produced a new extract for Hoffman-La Roche, for which he never published the extraction methods. With hindsight, this extract was probably much better than the first, if the convulsions that were reported were hypoglycaemic reactions (signs of a low blood-sugar level).

Many other researchers worked on the extraction of the internal secretion of the pancreas, both in Great Britain and the United States of America. Some, like Dewitt and Scott, used pancreatic duct-ligation to atrophy the acinar cells. This would remove the juice that would, as they believed, have destroyed the internal secretion. Others used alcohol (like Zuelzer) to remove the pancreatic juice. Among these were Knowlton and Starling, and Murlin and Kramer. Most of their results were irreproducible. Two Americans, Kleiner and Meltzer, did produce promising results of a decline in blood-sugar in depancreatised dogs, caused by administering extract of the dogs' own pancreases. Most of the control experiments were satisfactory. This work stopped abruptly in 1919 because Kleiner left the laboratory (Bliss, 1982).

There were several meetings among those scientists to discuss the prospects of

the research. Especially the intervention of a Scotsman, Macleod, who worked in Ohio, and who had been in contact with the researchers in Britain and seen their disappointing results, caused some of the above to stop their experiments.

On October 31, 1920, a fairly inexperienced general surgeon in London near Toronto (Canada), Banting, read an article on pancreatic duct-blockage. This gave him the idea of duct-ligation in order to isolate the internal secretion of the pancreas. This, as seen above, had been tried before, but the article did not mention it and Banting did not know. This is what he wrote in his note-book (Bliss, 1982):

"Diabetus
Ligate pancreatic ducts of dog. Keep dogs alive till acini degenerate leaving Islets.
Try to isolate the internal secretion of these to relieve glycosurea"

On November 8 he managed to arrange a meeting with Macleod, who worked in Toronto at that time. Although Macleod had discouraged several scientists in the field of pancreatic extracts, Banting's enthusiasm caused him to offer a laboratory and some animals over the next summer holiday period, and the help of a student, Best, as research assistant. In March 1921 Banting decided to take Macleod up on his offer, after which the first dogs were depancreatised on May 17. Macleod had advised to use Hédon's method of pancreatectomy leaving a small piece grafted under the skin to be removed later.

The first pancreatic extract was prepared and tested on July 30, with temporary success. The dog died a day later. The pancreas used had been removed seven weeks previously and left to degenerate. Another dog, brought back from a coma, also died within a day. In order to speed up, a full pancreatectomy was tried successfully on August 3, after which Hédon's procedure was not used again. Banting and Best named their preparation 'Isletin' in notes on the experiments on the dog that had the first total pancreatectomy.

Testing urine was sometimes difficult, because the volume of it decreased after injections of extract; in some cases urination ceased altogether. But during the second decade of the twentieth century methods of blood-sugar testing had been developed and improved, which were much more accurate than urine sampling. These tests made diabetes research more efficient and reliable.

Another time-saving idea was tried on August 17: extract of fresh, non-degenerated pancreas. Banting and Best failed to recognise the positive results and persisted with their faulty hypothesis that degeneration of the pancreas was necessary to obtain pure internal secretion. Boiling extract rendered it inactive, exhausting the pancreas' external secretion with secretin was too effective. Extracts prepared with secretin lowered blood-sugar quickly but caused profound shock. More and more control experiments were carried out, *e.g. in vitro* sugar-burning capacity and checking the activity of mixtures with trypsin. Macleod recognised it was this thorough testing that needed to be extended in order to make it impossible for critics and pessimists to deny the positive effects. The experiments would involve pre-injection blood tests, more frequent blood sampling after injection so as not to miss the effect, and establishing that it was a real blood-sugar lowering effect rather than dilution phenomena caused by the injections of reasonably large amounts of extracts.

In the first half of September 1921 different ways of injecting were tested. Rectal injections had no effect, while subsequent intravenous injection did prove effective. On September 17 injections were given subcutaneously for the first time, but the results were not satisfactory, and Banting and Best decided it was not worth trying again until they had trypsin-free extracts. By the end of September two more respected Toronto doctors/scientists, Starr and Henderson, had become involved in attempts to keep the promising research going by providing lab space and money.

Some time between October and December 1921 Best read a publication by Paulesco from July 1921. The blood-sugar data quoted by Paulesco were so different

from the generally accepted values for hyperglycaemia (probably due to different techniques used by Paulesco), that Best was not impressed and chose to ignore the information in the article.

In their first paper, describing work done up to November 10, which was published in February 1922 (Banting and Best, 1922), Banting and Best concluded that, although they had "always observed a distinct improvement in the clinical condition of diabetic dogs after administration of extract of degenerated pancreas", it was still too early for clinical trials. By then, they had started a so-called longevity experiment, keeping a pancreatectomised dog alive for as long as possible.

A visiting biochemist, Collip, became actively involved in designing better experiments. Banting discovered information about foetal pancreases, from which active extracts could be prepared without ligation or degeneration because of the relatively low content of acinar tissue; this meant that extract could now easily be produced in abundance from fresh, whole foetal pancreas. Time had come to try to capture "the active principle". New bacterial filtering methods produced more sterile extracts. Subcutaneous injections became possible, spreading action over a longer period, thus preventing shock. A new blood-sugar test was introduced, probably by Collip.

On November 23, Banting was injected subcutaneously with 1 1/2 cc of Berkefeld filtered extract. The group had become impatient, wanted to get into clinical testing. This extract did not seem to have any harmful effect, but blood-sugar was not measured.

One longevity experiment ended on December 2; the dog died after convulsions, due to anaphylactic shock. In retrospect, it could have been hypoglycaemic shock, since more extract seemed to make it worse. The next longevity experiment started four days later, in which the extraction was performed with alcohol instead of aqueous saline, which was easier to evaporate in order to concentrate the extract. This lead to the idea that the active principle could be extracted from adult pancreases

with alcohol too. Adult pancreases were available more cheaply than foetal pancreas. Although Banting and Best had started using alcohol and adult pancreases before Collip joined them, his experience and expertise were much needed assets in the group. He started using rabbits, and discovered that even healthy animals had a blood-sugar lowering response to extract. He also found that the residue and not the filtrate of a final filtering step contained the really powerful active principle. On December 20, 1921, Joe Gilchrist received tested, potent extract by mouth, with no benefit after a day. At that time it still was not firmly established that only injections would work.

Apart from testing urine and blood for sugar, Collip started testing the urine for ketones and measuring the liver's glycogen, which show whether liver function could be restored with the extracts. He also discovered hypoglycaemic shock, a state of apparent toxic reaction which could be relieved by administering glucose solution.

Because Collip was a much more self-sufficient, experienced and thorough researcher, his results were much better received than those produced by Banting and Best. Also, more and more people became involved in the experiments. This made Banting feel he had been overtaken, and that his idea had been taken out of his hands. Relationships within the group deteriorated quickly, and even became violent at times. Because of this animosity, Banting and Best decided to have a first official clinical test on January 11, 1922, on Leonard Thompson. The timing was wrong. It was still too early, and the extract did not perform as well as expected.

Meanwhile, Collip discovered the active principle could be purified to a certain extent by gradually precipitating other protein with increasing amounts of alcohol. Only at around 90% alcohol the active principle would precipitate, leaving it pure enough not to cause abscesses at injection sites. His extract underwent a first clinical test on January 23, also on Leonard Thompson. Two days later the group signed an agreement to work together as a group again rather than trying to compete with each other.

During February more clinical tests were carried out, all with favourable results. At the same time, Paulesco started clinical trials, independently and without knowledge of the work in Toronto. In April, the Latin-rooted name 'insulin' was proposed. The name 'insuline' had been suggested for the hypothetical internal pancreatic secretion twice before, in 1909 (de Meyer, 1909) and again in 1916 (Sharpey-Schafer, 1916), by two independent scientists, and unknown to the Toronto group.

On May 3, 1922, the discovery of insulin was officially announced to the medical world by Macleod.

1.4 The insulin molecule

After the actual isolation of insulin in 1922, it took another six years for Wintersteiner to establish that insulin is a protein (Wintersteiner *et al.*, 1928). And it was not until 1955 that the primary structure of insulin was elucidated by Sanger and co-workers. An account of this process can be found in the transcription of Sanger's Nobel Prize lecture (Sanger, 1959).

1.4.1 Finding the primary structure

Based on the knowledge of protein chemistry in general and the composition of insulin in particular, by 1943 Sanger started investigating the sequence of amino acids of insulin. Chibnall and his colleagues (Chibnall, 1942) had shown a high content of free α-amino groups, *i.e.* a relatively high number of N-terminal residues, one of which had been determined by Jensen and Evans (1935) to be phenylalanine. Until 1952 Sanger believed the molecular weight of insulin to be 12,000. In that year, Harfenist and Craig showed it to be around 6,000, using the method of partial substitution by 1-fluoro-2,4-dinitrobenzene (FDNB), separation of the reaction products and colorimetric analysis of the monosubstituted derivative for the dinitrophenyl

group.

For further study of the N-termini, Sanger developed the DNP-method (Sanger, 1945), which was later also used for many other proteins. The result was the discovery of two phenylalanine and two glycine termini, along with non-terminal ϵ-DNP-lysine. The assumed four polypeptide chains were thought to be linked by disulfide bridges due to the relative abundance of cysteines (du Vigneaud *et al.*, 1939). By oxidation with performic acid (Sanger, 1949a), the crosslinks were broken, resulting in two fractions, A and B. Fortuitously, the two types of amino acids that could have confounded this experiment by reacting with performic acid, methionine and tryptophan, are not present in insulin. Fraction A contained the smaller number (around 20) of residues, only 12 unique ones, of which none were basic. Four of them were cysteine. Fraction B had 30 residues, two of which were cysteine. It seemed there was only one type of glycine chain and one type of phenylalanine chain, which was confirmed in 1949 (Sanger, 1949b). Mild acid hydrolysis of the DNP-derivatives of the fractions made it possible to study the N-terminal sequences, resulting in Phe-Val-Asp-Glu (fraction B) and Gly-Ile-Val-Glu-Glu (fraction A).

This meant there were only two types of chain, and not four different ones, so the 12,000 molecular weight insulin was built up of two identical halves. Or, alternatively, the actual molecular weight was 6,000. During 1950 Sanger and Tuppy managed to sequence the whole of fraction B (Sanger and Tuppy, 1951a). Separation of the complex partial hydrolyzate by various means resulted in simpler mixtures of which direct analysis was possible. They ended up with a puzzle of about 45 peptides of varying lengths from which 5 sequences could be deduced:

Phe-Val-Asp-Glu-His-Leu-CySO_3H-Gly (N-terminal sequence)	(1)
Gly-Glu-Arg-Gly	(2)
Thr-Pro-Lys-Ala	(3)
Tyr-Leu-Val-CySO_3H-Gly	(4)
Ser-His-Leu-Val-Glu-Ala	(5)

Four amino acids were still missing, and it was impossible to establish how the sequences would be joined together. The use of proteolytic enzymes (pepsin, trypsin and chymotrypsin) (Sanger and Tuppy, 1951b) as hydrolytic agents instead of acid, solved this problem by producing different peptides, which also included the missing residues. Determination of the sequence of fraction A was completed by 1953 (Sanger and Thompson, 1953). This was more difficult because of the specific amino acid content and the lower susceptibility to enzymatic hydrolysis. Ionophoresis at around pH 3 was needed to separate the problematic cysteic acid peptides.

By that time, it was firmly established that the molecular weight of insulin was, in fact, 6,000. So the only remaining question, that of the number and nature of the disulfide bridges, was reduced to finding two between chains A and B, and one intrachain bridge in chain A. In order to do this, unoxidised insulin was subjected to hydrolysis to isolate peptides with intact cystines. Acid hydrolysis was not suitable because of disulfide rearrangement reactions, but enzymatic, neutral and alkaline hydrolysis all proved useful. Thus the complete sequence of insulin was deduced (see figure 1.1).

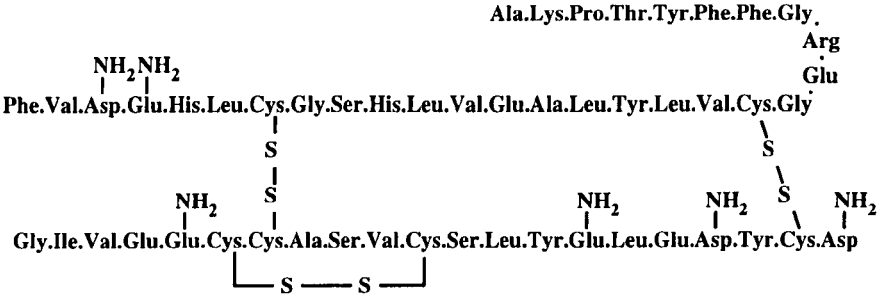

Figure 1.1: Sequence of insulin. Sanger used bovine insulin in his studies. Note the specific way of depicting amides. During strong acid hydrolysis the amide groups of some glutamines and asparagines are also hydrolysed. The real identity of these residues could be more appropriately determined by enzymatic hydrolysis (Sanger *et al.*, 1955).

Human insulin differs from bovine insulin at positions A8, A10 and B30, where

it has threonine, isoleucine and threonine, respectively. The molecular weight of human insulin is 5,808 (Hansen and Brange, 1987). The net charge of insulin is zero at pH 5.5, in good agreement with the isoelectric pH of 5.3-5.35 as originally determined by Wintersteiner and Abramson in 1933.

1.4.2 The three-dimensional structure of insulin

Insulin had been crystallised for the first time by Abel (1926). Over the years, the crystallisation method for these rhombohedral crystals was standardised. The discovery of Scott (1934) of the need for zinc in the crystals, was a big step forward in this. In 1935 Crowfoot received a first sample of finely crystalline insulin. She recrystallised the material according to Scott's method and took the first X-ray photographs, which were published in the same year (Crowfoot, 1935). They were not the first X-ray photographs of protein crystals ever, for pepsin had been used before (Bernal and Crowfoot, 1934). The first crystallisations were approached as an organic chemist would, so the crystals were dried with alcohol. Soon, however, the positive influence of mother liquor was an established fact, and X-ray measurements of wet insulin crystals were published (Crowfoot and Riley, 1939).

The interpretation of the X-ray patterns had started with the use of Patterson's ideas on the determination of the components of interatomic distances in crystals (Patterson, 1935). With the simplifications introduced by calculating Harker sections only (Harker, 1936), the maps showed, in the case of insulin, strong features at 10 and 22 Å interatomic distances. When the first amino acid crystal structure, glycine, was solved in 1939 through the application of the Patterson synthesis (Albrecht and Corey, 1939), solving protein structures with the same methods still seemed "madness", according to Patterson himself, as mentioned by Crowfoot Hodgkin in 1968 (Hodgkin and Riley, 1968). The idea of replacing zinc isomorphously with heavier ions had taken root almost as soon as the first photos had been taken. However, the intensity changes introduced by cadmium were too small

to be of any help. Iodination experiments were not followed through, probably partly because of the war, and because of the lack of experience with protein structure analysis in general. Until around 1950 most of the interpretation consisted in building models, almost by guessing and then checking whether they could produce the diffraction patterns observed. All the proposed models exhibited the correct symmetry, namely 32 (Hodgkin and Riley, 1968), but nobody recognised the possibility of calculating the correct size of the insulin molecule and the contents of the asymmetric unit then. The idea of utilising multiple crystal forms in structure determination was described by Crowfoot in 1938. Unfortunately, the many different shapes of crystals observed all turned out to be the same crystal form.

It was not until the primary structure of insulin became available in 1955 (Sanger *et al.*, 1955) that Hodgkin's group in Oxford started devoting most of their time to the problem of solving the three-dimensional structure of insulin. Schlichtkrull's methods of crystallisation (Schlichtkrull, 1958) and his investigations into the exact amount of zinc in those crystals, along with earlier experiments with cadmium insulin crystallisation according to Scott, paved the way for the production of lead insulin. It proved possible to remove zinc by soaking in EDTA and then replace the zinc with lead. X-ray measurements were taken of the isomorphous series of zinc-free, zinc, cadmium and lead insulin (Hodgkin and Riley, 1968). Especially the differences between metal-free insulin and lead insulin were large, but the feasibility of solving the structure with those alone, was still in doubt at that time.

However, it was not long at all until enough isomorphous heavy atom derivatives were found to solve the structure to a resolution of 2.8Å. Five derivatives were used for the phase determination:

- zinc-free insulin acetate, 0.01 M lead
- zinc insulin acetate, 0.1 M lead

- zinc insulin citrate, 0.02 M uranyl fluoride
- zinc insulin acetate, 0.01 M uranyl acetate
- zinc insulin citrate, mercuribenzaldehyde, saturated solution

Anomalous dispersion measurements were also taken. So late on Sunday night, August 3, 1969, insulin was solved (for a facsimile of the original note announcing the solution of the insulin structure, see Dodson, Glusker and Sayre (1981)), and the structure of rhombohedral 2-zinc insulin to a resolution of 2.8Å was published on November 1, 1969 (Adams *et al.*, 1969). The peptide chain, the disulfides and the aromatic residues were well defined. A few regions, especially on the outside of the molecule, were not completely clear.

After the determination of the primary structure of insulin by Sanger *et al.* in 1955, several groups in China started the chemical synthesis of insulin in 1958. They succeeded in 1965 (Kung *et al.*, 1965), after which some pressure was applied to determine the three-dimensional structure. The Cultural Revolution delayed the start until the beginning of 1967 (Tang, 1981). The structure of insulin to 2.5Å resolution was published in 1971 (Peking Insulin Structure Group, 1971). In 1972, when the Oxford group had extended their calculations to include data to 1.9Å resolution, Hodgkin visited China and it was decided the two groups would carry out further refinement separately and simultaneously. This resulted in publications by the Beijing group of the structure at 1.8Å resolution (Peking Insulin Structure Group, 1974) and at 1.2Å (Chang *et al.*, 1986), and at 1.5Å resolution by the Oxford group (Baker *et al.*, 1988). Meanwhile, both groups started studying different insulin species, for example insulin shortened at the B chain C-terminus (Peking Insulin Structure Group, 1976), 4-zinc insulin (Bentley *et al.*, 1976), insulin crystallised in cubic form (Dodson *et al.*, 1978) and insulin from hagfish (Cutfield *et al.*, 1979).

1.4.3 Biosynthesis of insulin

The biosynthesis of insulin starts in the nucleus of the B-cells of the pancreas. Most species have only a single insulin gene (Bell *et al.*, 1980). The DNA contains two intervening sequences, one in the 5' part of the mRNA that remains untranslated, and one almost in the middle of the sequence for the C-peptide region of proinsulin, 179 and 786 base pairs respectively. After transcription the mRNA ends up in the cytoplasm of the cell. It is thought (Steiner, 1983) that translation of the mRNA into the 110 amino acid preproinsulin (species: rat) starts while the ribosome is free in the cytosol. The signal sequence of the preproinsulin anchors the ribosome to the membrane of the rough endoplasmic reticulum (RER), after which the protein is translocated into the ER lumen. The 59 "extra" amino acids in the preproinsulin appear to contain all the information for accurate processing and secretion of the mature hormone. The prepeptide has many hydrophobic side chains and is exactly long enough to span the RER membrane so that the prohormone can start folding as soon as its N-terminus enters the ER lumen. There is evidence that this anchoring function of the prepeptide enhances the efficiency of protein folding. However, the prepeptide will be cleaved off by signal peptidase on the inner surface of the membrane before the whole of the peptide chain has entered the lumen. This is necessary for the correct bridging of the last two cysteines, one of which is the last-but-one in the nascent chain.

The length and not the amino acid content of the connecting peptide in proinsulin (26-35 residues of greatly varying sequence for all known species) appears its most important feature. Its function in folding and disulfide formation, *i.e.* maintaining the correct spatial separation of B30 and A1, could just as easily be exerted by a much shorter peptide. The reason for its excessive length may be that the prohormone chain needs to span the distance from the interface between the small and large ribosomal subunits (where translation takes place) to the inside of the ER, which is longer than 51 amino acids plus the necessary distance between B30 and

A1 in the mature hormone (Steiner, 1983). This is an example of the 'minimum length hypothesis' (Steiner, 1981). After folding and disulfide bridging, proinsulin is transferred to the Golgi apparatus. There the proteolysis of the prohormone starts and the mature hormone is concentrated, sorted and packed into secretory granules, ready for extracellular release.

1.4.4 Insulin function

The actions of insulin have been known for quite some time (Steiner, 1977):

- membrane transport of glucose, amino acids and certain ions;
- increased storage of glycogen;
- formation of triglycerides;
- stimulation of DNA, RNA and protein synthesis.

Three other peptide hormones are produced in the islets of Langerhans in the pancreas:

1. glucagon, consisting of 29 amino acids, in the A cells;
2. somatostatin, a cyclic 14 amino acid polypeptide, in the D cells;
3. pancreatic polypeptide, 36 amino acids with an amide C terminus, in the PP cells.

Glucagon antagonises most of insulin's actions, while stimulating insulin secretion. Somatostatin inhibits the three other islet hormones and a range of hormones from different origins. Pancreatic polypeptide inhibits pancreatic secretion altogether (Johnston *et al.*, 1988).

Once in the blood, insulin controls glucose homeostasis by stimulating the uptake of glucose into skeletal muscle and, to a lesser extent, into liver and adipose tissue. In muscle and adipocytes this uptake is mediated by the so-called

insulin-sensitive glucose transporter GLUT-4, a process that is not yet understood. Other processes in the regulation of glucose homeostasis are: alterations in glycogen metabolism in muscle and liver and decreased gluconeogenesis in the liver. The enzymes involved in the insulin-regulated processes of glucose metabolism appear to be regulated by (de)phosphorylation of serine and/or threonine residues (Lee and Pilch, 1994).

All known actions of insulin are initiated at the plasma membrane by insulin receptors responding to ligand binding. A schematic view of the insulin receptor can be seen in figure 1.2.

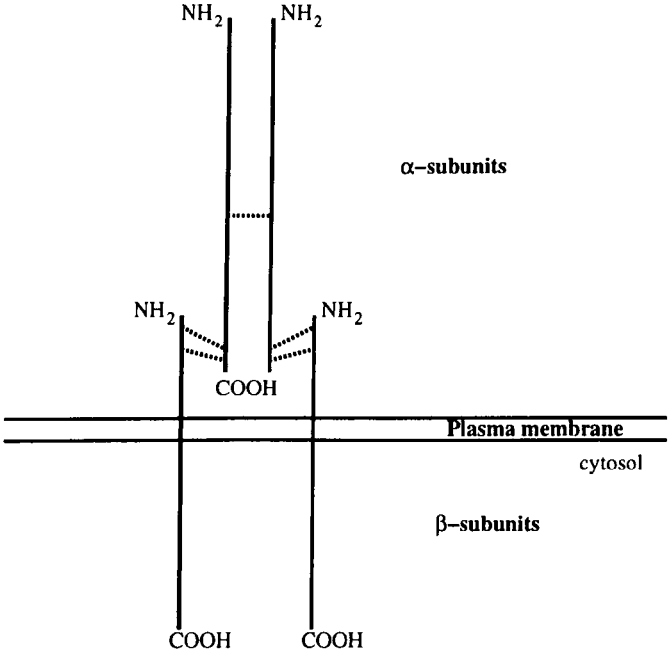

Figure 1.2: Schematic view of the insulin receptor. Reproduced from Lee and Pilch (1994)

The α-subunit of 723 amino acids contains the site or sites for insulin binding. The 620 amino acid β-subunit is built up of three regions: the extracellular, transmembrane and cytosolic domains. Both subunits are glycosylated, resulting in ap-

proximate molecular masses of 130 kDa and 95kDa respectively (Lee and Pilch, 1994). The insulin holoreceptor is composed of two α-subunits and two β-subunits covalently linked by disulfide bridges to form a functional dimeric protein complex. The receptor gene is translated into a single $\alpha\beta$-product of which two are linked together *via* disulfide bridges before being processed into separate α- and β-subunits. The covalent linkage of the two $\alpha\beta$-heterodimers is unusual among the receptor/tyrosine kinase family of which the insulin receptor is a member. The insulin holoreceptor can, however, under mild conditions be reduced to $\alpha\beta$-heterodimers which are functional monomeric receptor species with a lower ligand binding affinity than the holoreceptor. The exact position of the disulfide bridges both between α- and β-subunits and between $\alpha\beta$-heterodimers are not known; different studies are contradictory (Lee and Pilch, 1994).

Ligand-receptor contact occurs within the α-subunit. The exact regions of contact remain incompletely defined despite a lot of research effort. Various truncated versions of the receptor have been studied for insulin binding and it seems that only those with an intact α-subunit are capable of binding the ligand, leading to the conclusion that the entire α-subunit is important in some way for insulin binding. At physiologically relevant hormone levels there seems to be a stoichiometry of one insulin molecule per holoreceptor, with negative cooperativity for binding of a second ligand (Lee and Pilch, 1994). Thus insulin somehow breaks the symmetry of the receptor upon binding. Lee and Pilch propose a model for insulin binding to the receptor (see figure 1.3*a*) where at physiological concentrations insulin makes contact with two partial binding sites on separate halves of the receptor, resulting in high-affinity binding. It can be seen from the model that at high concentrations a second insulin molecule could potentially make contact with only one of the partial binding sites, resulting in low-affinity binding (figure 1.3*b*).

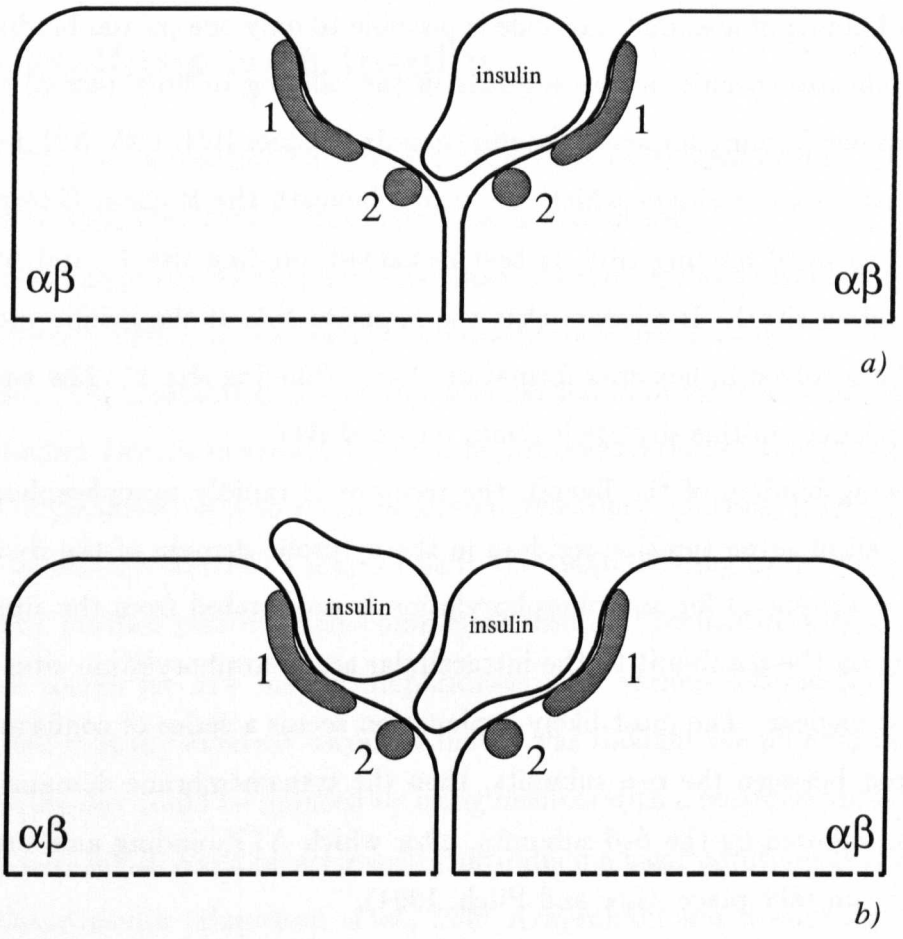

Figure 1.3: Schematic view of insulin binding as proposed by Lee and Pilch.
a) One insulin molecule per holoreceptor at physiological concentrations;
b) A second insulin molecule may bind, less strongly, at high concentrations.

In essence, the model proposed by Lee and Pilch is the same as that proposed by Schäffer (1994), with high-affinity binding of the first insulin molecule to two different partial binding sites on the two α-subunits of the receptor, after which low-affinity binding of a second molecule is possible to only one partial binding site. Schäffer implicates specific insulin residues in the binding to both partial binding sites. The dimer-forming surface of insulin (mainly residues B24, B25, A21 and B12, and potentially some residues which are buried beneath the B chain C-terminus), named the 'classical binding site', is best re-named 'binding site 1', and binds to receptor binding site 1. It appears that the opposite side of the insulin molecule, known to be involved in hexamer formation, forms 'binding site 2'. The two most important residues in this site are leucines A13 and B17.

Following binding of the ligand, the receptor is rapidly autophosphorylated on some or all of seven tyrosine residues in the cytosolic domain of the β-subunit. Exactly how the signal for autophosphorylation is propagated from the site of ligand binding on the α-subunit to the intracellular autophosphorylation sites on the β-subunit is unclear. The most likely explanation seems a series of conformational changes, first between the α-α-subunits, then the transmembrane domains of the β-subunits, followed by the β-β-subunits, after which ATP-binding and autophosphorylation can take place (Lee and Pilch, 1994).

The insulin receptor family primarily regulates nutritional metabolic pathways, whereas all other receptor/tyrosine kinases mainly regulate cell growth and differentiation. The same subdivision exists in that the insulin receptor family has covalent links between the $\alpha\beta$- heterodimers, and again in that these receptors do not form direct complexes with substrates and/or effector molecules after autophosphorylation. Instead, the insulin receptor, upon autophosphorylation of at least the tri-tyrosine subdomain, acquires exogenous kinase activity, phosphorylating its principal substrate: insulin receptor substrate 1 (IRS1). IRS1, in turn, binds effector molecules which are responsible for the actual processes of glucose transport and

metabolism.

1.5 Research into modern treatment of Diabetes Mellitus with insulin

Since 1922, a lot of research has gone into the improvement of insulin therapy, both in terms of improved insulin preparations and ease of use.

Achieving purity was the first challenge. Abel discovered insulin could be crystallised (Abel, 1926), which became standard procedure in insulin purification only after Scott (1934) established that zinc was needed in order to crystallise insulin in rhombohedral form, a discovery inspired by his observation of zinc in the pancreas. Another crystallisation step reduced allergic reactions (Jorpes, 1949). Chromatographic techniques started to play a role in the 1960s, leading to the first chromatographically purified insulin, Monocomponent insulin (Schlichtkrull *et al.*, 1970).

The search for new insulin preparations with various desired properties was approached in many different ways. At first it was thought the number of injections needed every day could be reduced by using insulins with a retarded uptake from the injection site, which could be achieved by introducing basic additives as in protamine and isophane insulin (Hagedorn *et al.*, 1936, Krayenbühl and Rosenberg, 1946). Addition of compounds like surfen or globin to acid insulin solutions, which produce heavily insoluble complexes upon neutralisation in tissue fluids, and complex formation of zinc with neutral insulin suspensions, had the same effect. However, the need for strict metabolic control in prevention of long-term complications called for the reinstatement of multiple injections, along with the development of rapid-acting insulins for the relief of the glucose-surge just after meal times, and mixtures of rapid-acting and intermediate-to-long-acting insulins. Developments in this field include Rapitard and Actrapid (Schlichtkrull, 1959, Schlichtkrull *et al.*, 1961).

The duration of action could also be influenced by the physical state and

size of the insulin particles. Ultralente (Hallas-Möller, 1956) is an example of a long-acting crystalline insulin preparation. Initially, only bovine insulin was used for crystalline preparations, having a slightly longer action than porcine insulin in crystalline form. Gradually, longer acting porcine preparations became available. Then, in 1979, recombinant DNA techniques made it possible to produce human insulin in *Escherichia coli* (Goeddel *et al.*, 1979). Amazingly, this was done without knowledge of the nucleotide sequence of the human insulin gene. Crea *et al.* (1978) had chemically synthesised two separate genes for chain A and B (77 and 104 base pairs for 21 and 30 amino acids, respectively, plus start and stop codons and restriction site bases) following nucleotide sequences that had been designed from the amino acid sequences. By designing the nucleotide sequences the way they did, there was no need to produce every possible trinucleotide separately. They produced 29 oligonucleotides, made from carefully chosen di-, tri- and tetranucleotide building blocks. Goeddel *et al.* (1979) describe the actual assembly of the genes, the subsequent construction of the plasmid, and the expression and characterisation of the product. They show that the amino acid content of the product is indistinguishable from that of porcine insulin. It took another year before the actual nucleotide sequence of the human insulin gene was published by Bell and co-workers (1980). Only 21 out of 51 of the codons used by Crea *et al.* turned out to be the same in the correct DNA-sequence.

Production of human insulin could also be achieved by conversion of porcine insulin (Markussen, 1982) or biosynthesis in *Saccharomyces cerevisiae* rather than *E. coli* (Markussen *et al.*, 1986). Over the years, the production of many therapeutic insulins has involved chemical alteration of the molecule. Nowadays, modified insulins can be synthesised by mutation of the genes used in *E. coli* or *S. cerevisiae*, thus facilitating structural and functional studies. But apart from the scientific possibilities opened up by the availability of recombinant insulin, there was a pressing need for a new source of the protein, because the demand for insulin for therapy

was outgrowing the supply of slaughter pancreases for the isolation of insulin. Even now (Kott, 1996) the major insulin manufacturers are not able to provide insulin for every area of the world, and cheaper beef and pork insulin is still produced, especially for third world countries.

Because of the varying needs of diabetic patients, insulin has to be available in multidose quantities. This requires the addition of antimicrobial preservatives. Banting and Best used tricresol to that effect, and to this day phenol and derivatives like *m*-cresol and methylparaben are used throughout the range of therapeutic agents. Other additives include sodium chloride or glycerol as isotonic agents, and certain buffers (Brange, 1987).

Alongside research into better and purer preparations, came the development of techniques of administering them. At the time of Banting and Best it was already established that oral therapy did not have any desired effect, and that subcutaneous or intravenous injections were the only option. Many other routes of absorption were tested, including rectal administration and absorption by mucosae. Success was limited. Even with more modern technology of aerosol powder (Wigley *et al.*, 1971), surfactants (Hirai *et al.*, 1981), liposome-enclosure (Dapergolas and Gregoriadis, 1976) or polymer-crosslinking (Saffran *et al.*, 1986) the desired efficiency and bioreactivity has not been achieved. The subcutaneous implantation of vinyl-ethylene copolymer pellets (Creque *et al.*, 1980, Brown *et al.*, 1986) or biodegradable insulin-albumin microbeads (Goosen *et al.*, 1983), releasing insulin slowly and constantly over a longer period of time, seems more promising. Most recently, reports have appeared on glucose-responsive insulin release from certain polymeric systems (Shiino *et al.*, 1995, Valuev, 1995). However, the only techniques in actual clinical use are based on injections. Hospitals operate systems of continuous infusion, either according to continuously measured glucose concentrations (Albisser *et al.*, 1974, Pfeiffer *et al.*, 1974) or a pre-programmed schedule (Slama *et al.*, 1974). These insulin pumps are not yet available to the general public,

although additional research is done on implantable pumps (Buchwald *et al.*, 1981). Today's most patient-friendly portable insulin delivery is the NovoPen, which has reduced injections to just pressing a button. Most diabetics seem to prefer this to other available options (Jefferson *et al.*, 1985, Walters *et al.*, 1985).

1.6 Classification of Diabetes Mellitus

In 1979, the National Diabetes Data Group formally classified Diabetes Mellitus and other categories of glucose intolerance as follows:

- Type I, insulin-dependent Diabetes Mellitus. Since this type usually occurs in juveniles, it was previously called juvenile diabetes. It can, however, start at any age. Patients usually present with easily recognisable symptoms, so diagnosis is not difficult. Genetic determinants seem important for the onset in most patients, with environmental factors a close second. Abnormal immune responses (*e.g.* in normal childhood diseases like mumps) and autoimmunity (see Bottazzo, 1993) are also thought to play an aetiologic role.

- Type II, noninsulin-dependent Diabetes Mellitus. This type can become recognisable at any age, although it can be asymptomatic for years and thus usually presents in patients over 40 years of age. Often only the complications seen after years of having diabetes, like neuropathy and cataracts, cause a diagnosis to be made. Occasionally people are diagnosed as a direct result of population studies, *e.g.* Mooy *et al.*, 1995 and Beks *et al.*, 1995, a study in Hoorn, The Netherlands. This research into the prevalence of NIDDM and predictors for the development of the disease resulted in 106 newly diagnosed diabetics out of 2,484 participants between age 50 and 74. Not only age is an important factor in NIDDM. The genetic basis of NIDDM seems even stronger than that of IDDM, and it is aggravated by environmental factors. Moreover, 60 to 90% of all NIDDM patients in the Western world are obese, which should be seen

as an indicator for subclassification of the type II diabetic. Symptoms are usually (at least partly) alleviated by weight loss.

- Other types of diabetes. Diabetes forms part of certain other conditions and syndromes, whether obviously aetiologically related or not. This class can be subdivided according to known or suspected aetiological relationships, where diabetes may be secondary to

  1. Pancreatic disease (neonatal or later on in life);
  2. Hormonal abnormalities which may have either hypoinsulinaemia or hyperinsulinaemia as a consequence;
  3. The administration of certain hormones, drugs and chemical agents, of which oral contraceptives, tricyclic antidepressants and marijuana are but a few;
  4. Insulin receptor abnormalities, either in the number of receptors or their affinity for insulin, or even because of the presence of antibodies to receptors (with or without associated immune disorders);
  5. Certain genetic syndromes, *e.g.* metabolism disorders, insulin resistance, hereditary muscle disorders and cytogenic disorders like Down's syndrome;
  6. Other types, of which diabetes associated with malnourished populations is the most prominent example.

- Gestational Diabetes Mellitus. There are two ways in which Diabetes Mellitus and pregnancy can occur simultaneously. One is where a previously diagnosed diabetic woman becomes pregnant, which involves certain risks for both mother and child. The other is in Gestational Diabetes, when a pregnant women becomes diabetic at some stage during pregnancy, because of the pregnancy and most commonly only for the duration of the pregnancy. It is

easy to see the origin of problems in both cases, if one considers the fact that even in a normal pregnancy the third trimester is a permanent state of mild hypoglycaemia because all organs have to work harder than normal, some up to 50%. Sometimes the pancreas functions sub-optimally. This means that the pregnant woman does not produce enough insulin herself, and the foetus will start to produce insulin to attain an acceptable maternal insulin level. This means the production of more foetal urine into the amniotic fluid. At the same time the level of growth hormone in the foetus rises along with the foetal insulin level (Remmers, 1994). This causes macrosomia (birthweight > 4,000 grams), which may cause problems at birth. A Glucose Tolerance Test usually points out the need for diet or insulin therapy. Oral hypoglycaemic treatment can not be used by a pregnant diabetic (Drury, 1988). Even after insulin treatment for GDM, most pregnant women return to normal glucose tolerance after delivery. If this is not the case, they have acquired clinical diabetes. From the previously gestational diabetics, 40% acquire overt Diabetes Mellitus within 20 years (Coustan, 1993) so they should be kept under some sort of observation.

- Impaired Glucose Tolerance. For the diagnosis of IGT an oral glucose tolerance test is essential. The criteria for this classification lie between those for normal subjects and diabetics. Consequently, the seriousness of the disorder is intermediate between normal and diabetic, with some clinical complications being completely absent while others, especially cardiovascular abnormalities, commonly present. Thus, IGT may have prognostic implications that should not be overlooked, especially in seemingly healthy individuals. Like overt diabetes, IGT can be linked to numerous disorders and obesity. Patients with IGT do not necessarily proceed to develop clinical diabetes; many return to normal glucose tolerance for no apparent reason while others stay in the IGT class for many years.

Apart from the clinical classes mentioned above, there are two so-called statistical risk classes (NDDG, 1979):

1. Previous Abnormality of Glucose Tolerance. This class is restricted to those persons who have normal glucose tolerance but have previously demonstrated diabetic hyperglycaemia or IGT either spontaneously or in response to an identifiable stimulus. Re-classification of gestational diabetics, former obese diabetics who have normal glucose tolerance after weight loss, and temporarily hyperglycaemic patients (due to trauma or injury) into this class is a useful tool for facilitation of follow-up of such patients. The likelihood of such persons developing clinical diabetes (again) should be considered to be increased.

2. Potential Abnormality of Glucose Tolerance. Persons who have never exhibited abnormal glucose tolerance but who are at substantially increased risk for the development of diabetes should be classified as PotAGT. Certain risks for development of IDDM and NIDDM are well established, such as being a relative of an IDDM or NIDDM diabetic or belonging to certain ethnic or racial groups, although the degree of risk for any of the specific circumstances is much less clear.

1.7 Studying the three-dimensional structure of insulin

As described above, human insulin consists of 51 amino acids, divided into two chains, commonly labelled A and B, with 21 and 30 amino acids respectively. The chains are linked by three disulfide bridges, two forming interchain cystines at A7-B7 and A20-B19, and one forming an intrachain cystine at A6-A11. A piece of antiparallel β-sheet is formed upon dimerisation: residues B23 to B28 of one monomer lie antiparallel to the same stretch in the other monomer. There are two very small

α-helices in the A chain, and a three turn α-helix running from residues B9 to B19 is found in every insulin structure known so far.

Although insulin is only a small protein hormone with a characteristic three-dimensional structure, the sequence varieties in nature and the biological activities of these different forms *in vitro*, are quite diverse. This diversity provides a tool for research into more effective and efficient insulins for the treatment of Diabetes Mellitus. Solving modified insulin structures will help understand the differences in structure-function relations of all these insulins. It is also thought that understanding properties (*e.g.* conformational changes) in a small protein like insulin will contribute to the understanding of proteins in general.

Chapter 2

Protein crystallographic methods

2.1 Introduction to protein crystallography

A crystal is a solid with regular three-dimensional ordering. The regular repeat of the unique unit (unit cell) in a crystal can be represented by a regular repeat of points, the lattice, with the same geometry as the original unit cells. The symmetry of the lattice should reflect the symmetry of the crystal. This means that the symmetry most representative of the crystal does not always give rise to a primitive cell. A primitive unit cell can be described by a single representative point, commonly chosen to be the origin (0,0,0), from which the crystal can be built up by generating points with unit translations along the unit cell edges. Non-primitive lattices would have additional representative points translated 1/2 along appropriate cell edges.

The lattice, or lattice function $l(\vec{r})$, can be seen as a three-dimensional array of δ-functions corresponding to the lattice points. The unit cell function $u(\vec{r})$ describes the detailed information of the structure of the molecule(s) in the unit cell. The crystal structure function $c(\vec{u})$ is the convolution of the lattice function with the unit cell function (Sherwood, 1976):

$$c(\vec{u}) = l(\vec{r}) \otimes u(\vec{r}) \tag{2.1}$$

The unit cell contains one or more molecules built up of atoms. In a classical description, the charged particles in the atoms, namely the protons and electrons, interact with electromagnetic radiation like X-rays, producing the diffraction pattern observed when a crystal is put in an X-ray beam. The particles start to oscillate, becoming a source of radiation themselves, thus scattering the incoming wave. Because the intensity of scattered radiation is inversely proportional to the square of the mass of the scattering particle, electrons scatter X-rays around one million times more effectively than protons. It is therefore the electron cloud of atoms that is responsible for the abovementioned diffraction pattern. The scattering from the electron cloud is proportional to that of a free electron, assuming the electrons oscillate in phase. In other words, the scattering of a volume element $d\vec{r}$ at position $\vec{r}$ is proportional to the electron density $\rho(\vec{r})$ at $\vec{r}$. The use of the term electron "cloud" implies a finite size to the space the electrons around an atom may occupy. The approximation of an atom by a point-atom, useful though it may be in other cases, would in an X-ray experiment for a single atom give rise to equal scattering in all directions due to lack of interference between wavelets scattered by different volume elements $d\vec{r}$. However, the real dimensions of the electron cloud are comparable to the wavelength of the X-rays used, resulting in interference which causes scattering to fall off with increasing scattering angle θ. It can be shown that thermal motion, influencing the apparent size of the atom, causes the scattering to fall off more rapidly with θ. Atoms can be assumed to have spherical symmetry, so the scattering from an individual atom is a function which depends merely on θ. This function is known as the 'atomic form factor' or 'atomic scattering factor' and has been tabulated for all chemical elements (Wilson, 1992).

The diffraction pattern resulting from the scattering of X-rays by a distribution of electrons characterised by an electron density function $\rho(\vec{r})$ is the Fourier transform $\mathbf{F}(\Delta\vec{k})$

$$\mathbf{F}(\Delta\vec{k}) = \int_{all\ \vec{r}} f_e \rho(\vec{r}) e^{i\Delta\vec{k}\cdot\vec{r}} d\vec{r} \tag{2.2}$$

where f_e is the electronic scattering factor and $\Delta\vec{k}$ is the scattering vector, which can be expressed as $\Delta\vec{k} = \vec{k_d} - \vec{k_i}$. Here $\vec{k_i}$ is the wave vector of the incident beam and $\vec{k_d}$ the wave vector of the diffracted beam.

If the electron density $\rho(\vec{r})$ is expressed as a function of fractional coordinates x,y,z and only a unit cell is considered, for which f_e can be seen as a constant that can be taken up in **F**, one gets

$$\mathbf{F}(\Delta\vec{k}) = V \int_{x=0}^{x=1} \int_{y=0}^{y=1} \int_{z=0}^{z=1} \rho(x,\vec{y},z) e^{i\Delta\vec{k}\cdot(x\vec{a}+y\vec{b}+z\vec{c})} dxdydz \qquad (2.3)$$

In order to obtain a complete crystal, the unit cell is convoluted with the crystal lattice. The amplitude of the unit cell diffraction pattern, as described by equation 2.3, envelops the diffraction pattern of the crystal lattice, which is a series of sharp peaks forming the reciprocal lattice. The effect of this envelopment is the sampling of the unit cell diffraction function at points where the crystal lattice gives rise to a diffraction maximum. The diffraction pattern intensities, representing the reciprocal lattice weighted by the unit cell contents, can thus be described by means of delta functions (Sherwood, 1976):

$$|F(\Delta\vec{k})|^2 = [\sum_{h=-\infty}^{\infty} \delta(\Delta\vec{k}\cdot\vec{a}-2h\pi)]^2 [\sum_{k=-\infty}^{\infty} \delta(\Delta\vec{k}\cdot\vec{b}-2k\pi)]^2 [\sum_{l=-\infty}^{\infty} \delta(\Delta\vec{k}\cdot\vec{c}-2l\pi)]^2 \quad (2.4)$$

where $\vec{a}$, $\vec{b}$ and $\vec{c}$ are unit cell edge vectors, and h, k and l are integers. This function can only be non-zero at points where $\Delta\vec{k}\cdot\vec{a} = 2h\pi$, $\Delta\vec{k}\cdot\vec{b} = 2k\pi$ and $\Delta\vec{k}\cdot\vec{c} = 2l\pi$, known as the Laue equations. In other words, $\Delta\vec{k}$ is restricted to those values which correspond to the crystal lattice diffraction maxima, so $\Delta\vec{k}$ is proportional to the reciprocal lattice vector $\vec{G}_{hkl}$ as follows:

$$\Delta\vec{k} = 2\pi\vec{G}_{hkl} \qquad (2.5)$$

$\Delta\vec{k}$ forms an equilateral triangle with the abovementioned vectors $\vec{k_i}$ and $\vec{k_d}$

(see equation 2.2), where the angle between the incident and diffracted beam vectors is called the 'scattering angle' 2θ.

$$|\Delta\vec{k}| = 2ksin\theta \tag{2.6}$$

where k is the magnitude of both the incident and the diffracted beam vectors; k is also equal to $\frac{2\pi}{\lambda}$ because of wave theory. The magnitude of the reciprocal lattice vector $\vec{G}_{hkl}$ in equation 2.5 is inversely proportional to the spacing d_{hkl} between reciprocal lattice planes (hkl). Thus, following from equation 2.6, $\frac{1}{d_{hkl}} = \frac{2}{\lambda}sin\theta$, so

$$\lambda = 2d_{hkl}sin\theta \tag{2.7}$$

Equation 2.7 is known as Bragg's Law. The angle θ is called the 'Bragg angle', and is equal to half of the scattering angle. The wave vector triangle described above can be used in a geometrical interpretation of Bragg's Law: the Ewald sphere (see figure 2.1). The radius of the sphere is taken as $k/2\pi = 1/\lambda$.

From this construction it can be seen that, when the crystal rotates around C and its reciprocal lattice rotates around O, diffraction maxima only occur when wave vector $\Delta\vec{k}$ is an exact chord of the Ewald sphere, *i.e.* when a reciprocal lattice point passes through the Ewald sphere. This chord is the reciprocal lattice vector $\vec{G}_{hkl}$.

For $\vec{G}_{hkl} = h\vec{a}^* + k\vec{b}^* + l\vec{c}^*$ the expression $\Delta\vec{k} \cdot \vec{r}$ in equation 2.2 becomes $2\pi(hx + ky + lz)$. Equation 2.3 may now be rewritten as

$$\mathbf{F}_{hkl} = V\int_0^1\int_0^1\int_0^1 \rho(x,y,z)e^{2\pi i(hx+ky+lz)}dxdydz \tag{2.8}$$

which is known as the 'structure factor' of the diffraction maximum at reciprocal lattice point hkl. $\mathbf{F}_{hkl}$ is a complex quantity and can be expressed as $\mathbf{F}_{hkl}=|F_{hkl}|e^{i\alpha_{hkl}}$. The experimentally observable intensity is $I_{hkl}=|F_{hkl}|^2$.

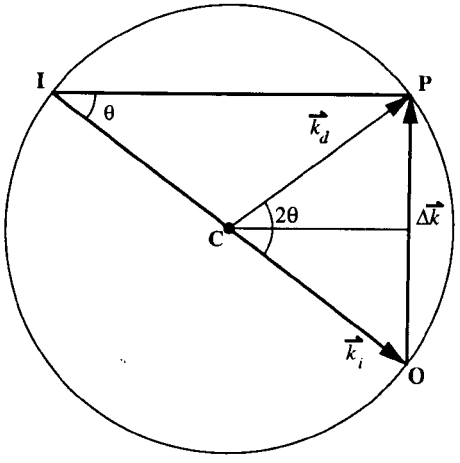

Figure 2.1: The Ewald sphere construction.
I = point where incident beam crosses the Ewald sphere; C = position of the crystal; O = exit point of direct beam, commonly chosen as the origin of reciprocal space; P = exit point of diffracted beam; $\vec{k}_i$ = incident wave vector; $\vec{k}_d$ = diffracted wave vector; $\Delta\vec{k}$ = reciprocal lattice vector; θ = Bragg angle; 2θ = scattering angle.

One of the more interesting properties of a diffraction experiment is that the diffraction pattern is the Fourier transform of the electron density function, which means that the electron density function $\rho(x, y, z)$ can be expressed as

$$\rho(x, y, z) = \frac{1}{V} \int_{all\ \Delta\vec{k}} \mathbf{F}_{hkl} e^{-2\pi i(hx+ky+lz)} d(\Delta\vec{k}) \tag{2.9}$$

Because the diffraction pattern is not a continuous function of $\Delta\vec{k}$ but a discrete set of values represented by $\mathbf{F}_{hkl}$ for each maximum, this expression reduces to

$$\rho(x, y, z) = \frac{1}{V} \sum_h \sum_k \sum_l \mathbf{F}_{hkl} e^{-2\pi i(hx+ky+lz)} \tag{2.10}$$

Assuming all $\mathbf{F}_{hkl}$ can be derived from the diffraction pattern, the electron density can be calculated at every point x,y,z by adding up the contributions from all the reflections hkl. The total density function $\rho(x, y, z)$ can be computed from

the local densities at every point x,y,z. The computation of the electron density from structure factors is called 'Fourier synthesis'.

Unfortunately, experimental data does not provide amplitudes $\mathbf{F}_{hkl}$, but intensities I_{hkl}, where $I_{hkl}=|F_{hkl}|^2$. For a Fourier synthesis, the complete information of $\mathbf{F}_{hkl}$, *i.e.* both $|F_{hkl}|$ and the phase α_{hkl}, are needed. Thus part of the information is lost, giving rise to what is called the Phase Problem.

2.2 Crystallisation

The crystallisation of proteins and protein-like substances has been a science for over a century. Van Deen (1864) believed that all naturally occurring organic substances can be crystallised when manipulated effectively. Schimper (1881) reported on the first active crystallisation attempt by Maschke in 1859, who evaporated the solution from a preparation of brazil nuts in order to obtain crystals.

Insulin exists in crystalline form in storage granules in the pancreas. Crystallisation of insulin was pioneered by Abel in 1925, after which Scott established the need for zinc for successful reproducible crystallisation (Scott, 1934). The first insulin crystals used in X-ray work were grown according to his methods (Crowfoot, 1935). Over the decades, the protocols for insulin crystallisation were standardised and extended to produce different crystal forms. Schlichtkrull (1958) had a major role in the standardisation of the crystallisation with zinc and other metals, ultimately leading to the successful crystallisation of suitable heavy atom derivatives for the solution of the structure of native 2Zn insulin in 1969 (Adams *et al.*, 1969). It became clear that a minimum amount of 2 Zn atoms per rhombohedral cell (trigonal setting) was needed, or 4 Zn atoms per cell in crystallisations with $\geq$ 6% halide. There is also a maximum amount of 6 Zn atoms per cell, beyond which no crystals will form or crystallisation is very slow (Schlichtkrull, 1958, Brange, 1987).

Brange (1987) shows a very elegant graph of the crystallisation conditions

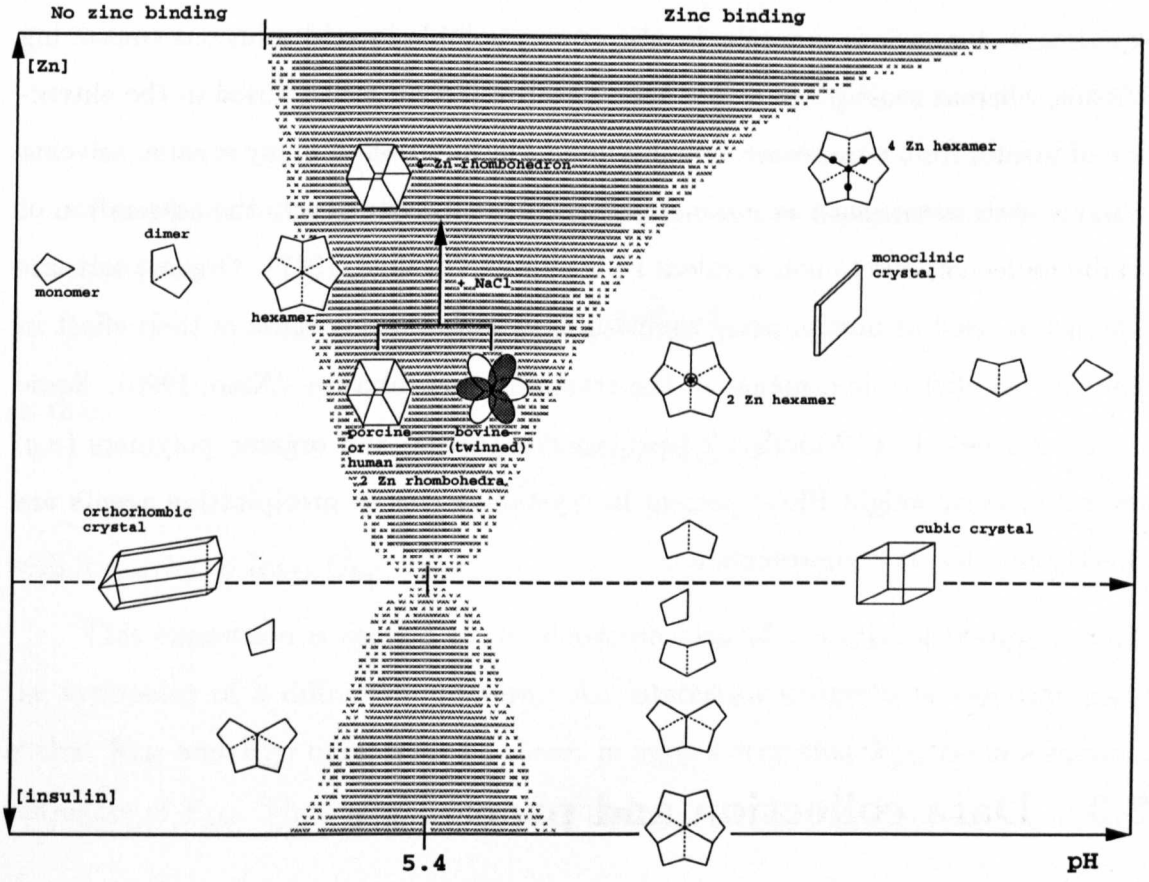

Figure 2.2: The crystallisation behaviour of insulin. Reproduced from Brange (1987)

for insulin in its various crystal forms (see figure 2.2). In general, crystallisations of modified insulins are carried out with the purpose of the modification in mind. Mutants which form supposedly more stable hexamers than native insulin, are crystallised following the 2Zn or 4Zn protocols described in table 3.2 in chapter 3. For even more stabilisation of the hexamer, phenol is added in some cases, according to the monoclinic insulin crystallisation protocol. In order to study metal binding, metal salts are added where appropriate, resulting in either aggregation to hexamers (for which metal ions are necessary in general) or dimers (which can be formed without metal ions, or may still be formed in their presence when the metal bind-

ing site is removed). Organic solvents like ethanol and acetone are used in insulin experiments for various reasons. Insulin is very soluble in acid solutions containing ethanol, whereas most proteins are not. This property is still utilised in the extraction of insulin from pancreases. Furthermore, the addition of many organic solvents miscible with water (such as ethanol and acetic acid) counteracts the association of insulin molecules due to non-covalent interactions (Brange, 1987). Organic solvents can also be used as precipitating agents in crystallisations because of their effect in reducing the dielectric constant of the crystallisation mixture (Xiao, 1990). Some organic solvents (*e.g.* 2-methyl-2,4-pentanediol, MPD) and organic polymers (*e.g.* lower molecular weight PEG) present in crystallisations as precipitating agents are directly suitable as cryoprotectants.

2.3 Data collection and processing

Equation 2.8 can be rewritten in a form more suitable for the calculation of structure factors from the contents of the unit cell, and for the discussion of the symmetry of a diffraction pattern (Sherwood, 1976). Introducing the atomic electron density function $\rho_j(\vec{r}-\vec{r}_j)$ for every atom j, based on a vector triangle $\vec{r}=\vec{r}_j+\vec{R}_j$ defining the position of an electron with respect to its atomic nucleus ($\vec{R}_j$) and the position of the atom in the unit cell ($\vec{r}_j$), equation 2.8 becomes

$$\mathbf{F}_{hkl} = \int_{unit\ cell} \sum_j \rho_j(\vec{R}_j) e^{2\pi i \vec{G}_{hkl}\cdot\vec{r}_j} e^{2\pi i \vec{G}_{hkl}\cdot\vec{R}_j} d\vec{R}_j \qquad (2.11)$$

assuming the positions of the atomic nuclei are constant so that $d\vec{R}_j$ is equal to $dxdydz$. Reversing the integral and the sum, factoring the constant $e^{2\pi i \vec{G}_{hkl}\cdot\vec{r}_j}$ out of the integral and changing the limits of the integral from unit cell to atom (re-building

the integral with the summation over all j atoms), $\mathbf{F}_{hkl}$ becomes

$$\mathbf{F}_{hkl} = \sum_j e^{2\pi i \vec{G}_{hkl}\cdot\vec{r}_j} \int_{atom} \rho_j(\vec{R}_j) e^{2\pi i \vec{G}_{hkl}\cdot\vec{R}_j} d\vec{R}_j \tag{2.12}$$

The aforementioned atomic scattering factor can now be reintroduced in mathematical form:

$$f_j = \int_{atom} \rho_j(\vec{R}_j) e^{2\pi i \vec{G}_{hkl}\cdot\vec{R}_j} d\vec{R}_j \tag{2.13}$$

so that

$$\mathbf{F}_{hkl} = \sum_j f_j e^{2\pi i(hx_j+ky_j+lz_j)} \tag{2.14}$$

with $hx_j + ky_j + lz_j = \vec{G}_{hkl}\cdot\vec{r}_j$.

This expression is central to the determination of the crystal symmetry and the symmetry of a diffraction pattern. An interesting property of equation 2.14 is that $\mathbf{F}_{hkl}$ and $\mathbf{F}_{\bar{h}\bar{k}\bar{l}}$ only differ in phase, in such a way that $\mathbf{F}_{\bar{h}\bar{k}\bar{l}}$ is the complex conjugate of $\mathbf{F}_{hkl}$. Their magnitudes are equal:

$$|F_{hkl}| = |F_{\bar{h}\bar{k}\bar{l}}| \tag{2.15}$$

which is known as Friedel's Law. Thus the observed intensities of reflections hkl and $\bar{h}\bar{k}\bar{l}$ are equal. This means that a weighted reciprocal lattice will always be centrosymmetrical in nature. Equation 2.14 can also be used in the determination of the presence of translational symmetry in the crystal. A simple example is reproduced from Sherwood (1976). In the case of a body-centred cubic lattice for metallic caesium with only two independent sites, (0,0,0) and (1/2,1/2,1/2), the structure factor only has two terms:

$$\mathbf{F}_{hkl} = f_j e^{2\pi i(h0+k0+l0)} + f_j e^{2\pi i(h1/2+k1/2+l1/2)} \tag{2.16}$$

Since the atomic contents of the two sites are equivalent, f_j is equal in both cases,

so

$$\mathbf{F}_{hkl} = f_j[1 + e^{i\pi(h+k+l)}] \tag{2.17}$$

Because h,k and l are integral numbers, $\mathbf{F}_{hkl}$ can only be 0, if $h + k + l$ is odd, or a finite value $2f_j$, if $h + k + l$ is even. Half of the diffraction maxima are missing, the so-called systematic absences. It can be shown that all translational symmetry gives rise to specific systematic absences. The systematic absences corresponding to lattice type unambiguously determine, given the crystal system, what is called the Bravais lattice, of which there are 14.

The reciprocal lattice, and thus a diffraction pattern, does not explicitly have translational symmetry. Therefore, the symmetry of a diffraction pattern must be that of one of the 32 point-groups. Only eleven of those have the, by Friedel's Law, required centrosymmetry, and are referred to as the Laue groups.

Buerger has calculated there are 122 different possible symmetry types for diffraction experiments, the diffraction symbols, one of which must fully describe the diffraction pattern. Fifty-eight of these diffraction symbols uniquely define a space group, the others give rise to ambiguity. Many of the latter 64 contain mirror planes and inversion centres, which makes them impossible for natural proteins. So usually in protein crystallography the space group of a crystal is reasonably easily determined from a diffraction pattern, with the help of data processing computer programs.

The theory of data processing is mainly concerned with correcting for the assumptions and approximations made above. What is actually measured in a diffraction experiment, is the total energy scattered by the crystal towards the diffraction maximum hkl. Sherwood (1976) calls this quantity E_{hkl}:

$$E_{hkl} = \frac{I_0}{\omega} K L_{hkl} p_{hkl} A'_{hkl} e_{hkl} |(F_{hkl})_T|^2 V \tag{2.18}$$

The proportionality with the intensity of the incident beam I_0 and the volume of the

crystal V is intuitively obvious. A'_{hkl} and e_{hkl} correct for absorption and extinction effects. ω is the angular velocity of the rotation of the crystal, which comes into play when considering the amount of time needed for a reciprocal lattice point to pass through the Ewald sphere. K represents a number of fundamental constants involved with the physics of the experiment (*e.g.* the mass of an electron and the wavelength of the radiation used). The polarisation factor p_{hkl} is specific for the type of incident beam used, with different values for *e.g.* randomly polarised and monochromated beams. The geometry of the diffraction experiment and the position of a diffraction maximum relative to O (see figure 2.1), have an effect on what is recorded when the reciprocal lattice point passes through the Ewald sphere, which is taken into account with the Lorentz factor:

$$L_{hkl} = \frac{1}{\lambda}\frac{\omega}{\nu_n} \tag{2.19}$$

where ν_n is the component of velocity ν normal to the Ewald sphere surface in the direction of C. ν is the linear velocity in the direction $\vec{PI}$ if the crystal is rotating with angular velocity ω around an axis through O perpendicular to the plane of the paper (see figure 2.1). λ is the wavelength of the X-ray experiment.

The effect of thermal motion is taken up in $(F_{hkl})_T$, which has $(f_j)_T = f_j e^{-B_j(sin^2\theta/\lambda^2)}$ as atomic scattering factor rather than just f_j, and thus falls off more rapidly with θ, as discussed in section 2.1. B_j is the so-called temperature factor.

In protein crystallography there is another major factor of influence in data collection: protein crystals suffer radiation damage. Performing the diffraction experiment at cryogenic temperatures will reduce this effect dramatically. Any remaining damage needs to be corrected for during data processing.

Apart from the desired discrete diffraction due to the crystal, there is continuous background scattering of X-rays due to scattering by the air and incoherent

scattering of the crystal (a quantum mechanical effect). The final number assigned to a diffraction maximum is obtained by integration, the total area of the spot, rather than the absolute peak height. It depends on numerous parameters of the specific diffraction experiment, and thus can not be determined on an absolute scale. After integration, equivalent reflections are scaled and merged, and standard deviations determined by estimation from the counting statistics, attempting to take into account systematic errors of the diffraction experiment. Because of counting errors, some of the resulting measured intensities will have a negative value, although a true intensity can, of course, not be negative (French and Wilson, 1978). If left negative, such measurements will pose problems upon taking the square root when calculating the structure factor modulus. Simply omitting them will bias the resulting structure, as will resetting them to zero without adjusting their standard deviations. French and Wilson have devised a way of treating negative intensity measurements more advantageously, based on Bayesian statistics. Their methods only have an effect on the smallest terms, without the discontinuity observed in other methods, and without resetting them to a constant value. This should result in least possible loss of information and least possible bias. Unfortunately, some data processing programs still insist on the conventional approach of resetting to zero or omitting.

2.4 Structure Determination

When a new structure can be assumed to be isomorphous with one that is already known, structure determination is simple: $(2F_{obs} - F_{calc})$ and $(F_{obs} - F_{calc})$ maps can be calculated with the observed magnitudes of $\mathbf{F}_{hkl}$ from the new structure, and the calculated magnitudes and phases of $\mathbf{F}_{hkl}$ from the presumed isomorphous known structure. These maps should show the small differences between the two structures well. Isomorphism can be assumed on the basis of a strong similarity between the

cell constants of two structures, and a low value for the isomorphous R-factor:

$$R_{iso} = \frac{\sum_{hkl} ||F_{PH}| - |F_P||}{\sum_{hkl} |F_P|} \tag{2.20}$$

In all other cases the crystallographer has to deal with the aforementioned Phase Problem. For small molecules this problem can be solved with several methods based on the Patterson function, or by so-called Direct Methods. Direct methods are based on statistical relationships between structure factors, and often involve probability functions which depend on the reciprocal of the number of atoms. Moreover, they usually require data to extend to atomic resolution. For proteins, with relatively large numbers of atoms and generally lower than atomic resolution data, these methods are in general not powerful enough to solve the structure. However, a lot of effort is applied to determine ever larger structures *ab initio*, with some successes for small proteins and peptides (see for example Sheldrick *et al.* 1993).

The Patterson function is defined as

$$P(u,v,w,) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}|^2 e^{-2\pi i(hu+kv+lw)} \tag{2.21}$$

which is the inverse Fourier transform of the relative intensities $|F_{hkl}|^2$ which are the experimentally measurable quantities. Because of the symmetry of direct and inverse Fourier transforms, it is valid to say the Patterson function is simply the Fourier transform of $|F_{hkl}|^2$. Invoking some complex number theory, the convolution theorem and the notion that the electron density is the Fourier transform of the structure factor, it can be proved that

$$P(u,v,w) = V[\rho(x,y,z) \otimes \rho(-x,-y,-z)] \tag{2.22}$$

Or, in words, the Patterson function is the convolution of the electron density with its centrosymmetric image, scaled by the volume of the unit cell. (u,v,w) represents

a set of vectors between all pairs of atoms in the unit cell. The peaks in the Patterson function occurring at (u, v, w) have the strength of the product of the strengths at either side of the vector, which are directly linked to the number of electrons of the atoms in the atom pair, apart from scale factors. For a structure with N atoms there are N(N-1) non-origin peaks. Thus the Patterson map of a protein would contain a large number of peaks, many of which coincide or overlap, since the width of a Patterson peak is the sum of the widths of the appropriate two peaks in the corresponding electron density maps. This generally makes the direct determination of a protein structure from a Patterson map impossible, although there are exceptions (see for example Frazão *et al.*, 1995). The protein crystallographer therefore has to resort to one of two other methods to overcome the Phase Problem: isomorphous replacement or molecular replacement. Molecular replacement can be used when the structure of a similar protein is known, *e.g.* a native protein in the case of a mutant, the same protein from a different species, a model from NMR studies, or an otherwise homologous protein which might serve as a model for the protein under study. In the case of a completely "new" protein for which no related structure is known, isomorphous replacement is the only available option. It is by this method that the very first insulin structure was determined (Adams *et al.*, 1969).

2.4.1 Isomorphous replacement

The isomorphous replacement method is based on the heavy atom technique as developed for small molecules. The idea is that the diffraction pattern and phases are dominated by the contribution of one or a few heavy atoms if their atomic number is sufficiently large relative to that of the other atoms present. The position of the heavy atoms can be determined from a Patterson map, after which an electron density map can be phased on those atoms only. This map should reveal the positions of the other atoms. If no heavy atom is present, a suitable atom might be replaced by a heavy atom. If this heavy atom derivative can be crystallised isomorphously,

the heavy atom method will facilitate the solution of both the native and modified structures. In proteins the production of a heavy atom derivative is often reduced to the soaking in of suitable heavy atoms into the solvent channels. When the procedure is effective, the structure of the protein is not disturbed and the heavy atoms bind sufficiently to certain groups. The small changes in the diffraction pattern can then be directly attributed to the heavy atoms. If $\mathbf{F}_P$ is the structure factor for the native protein and $\mathbf{F}_{PH}$ that for the derivative, the contribution of the heavy atoms is

$$\mathbf{F}_H = \mathbf{F}_{PH} - \mathbf{F}_P \tag{2.23}$$

Although only the magnitudes $|F_P|$ and $|F_{PH}|$ can be measured, the vector $\mathbf{F}_H$ can be calculated once the positions of the heavy atoms have been determined. The interpretation of the Patterson function can be made easier by taking $(|F_{PH}| - |F_P|)^2 = |F_H|^2 cos^2(\alpha_{PH} - \alpha_H)$ as coefficients. This is called isomorphous-difference Patterson Synthesis and in the resulting Patterson map all the available information which the native protein and the derivative have in common, has cancelled out, leaving a clearer picture of the heavy atom information, although some noise has been introduced. This noise stems from the cosine term in the derivation of the coefficients: $\frac{1}{2}|F_H|^2 + \frac{1}{2}|F_H|^2 cos2(\alpha_{PH} - \alpha_H)$. Since α_{PH} and α_H are not correlated, this term only contributes noise.

In centric zones the structure factors are real quantities, so $|F_H| = |F_{PH}| \pm |F_P|$. Usually, because $|F_H|$ is relatively small, the difference is the correct estimate. Only when $|F_{PH}|$ and $|F_P|$ are both also small, the sum can be correct. In acentric cases the Phase Problem is more complicated. It can be solved with the aid of a phasing diagram as devised by Harker (1956). Figure 2.3 shows a phasing diagram for one isomorphous derivative. It can be seen that the phasing circles for the native protein and that for the derivative intersect at two points, giving rise to two phase triangles fulfilling equation 2.23. The Phase Problem is not solved unambiguously.

This can be remedied by adding another derivative circle as in figure 2.4.

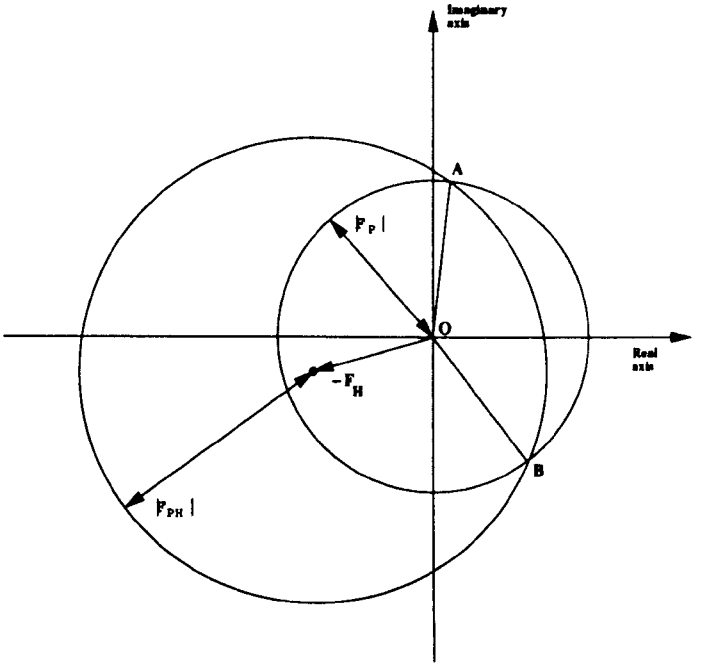

Figure 2.3: Phasing diagram for Single Isomorphous Replacement. Two circles are drawn, one with radius $|F_P|$ centred on the origin, and one with radius $|F_{PH}|$ about the end of the vector $-\mathbf{F}_H$. The two points of intersection A and B of the two circles correspond to two possible values of the protein phase.

So long as the heavy atoms do not occupy the same sites, *i.e.* the derivatives are sufficiently different, and the native protein and the derivatives are sufficiently isomorphous, so that the phasing circles are well-defined, the ambiguity is resolved. The correct solution is the point where the three circles intersect.

Another method of resolving the ambiguity is to take into account that the electrons in the crystal are not really free; atoms scatter X-rays as if they contain electronic dipole oscillators. This phenomenon, called anomalous scattering, will be especially noticeable if the X-ray wavelength is just shorter than that which corresponds to an absorption edge of the atom in question, so that the electrons in the atom can change quantum state. The atomic scattering factor depends on the

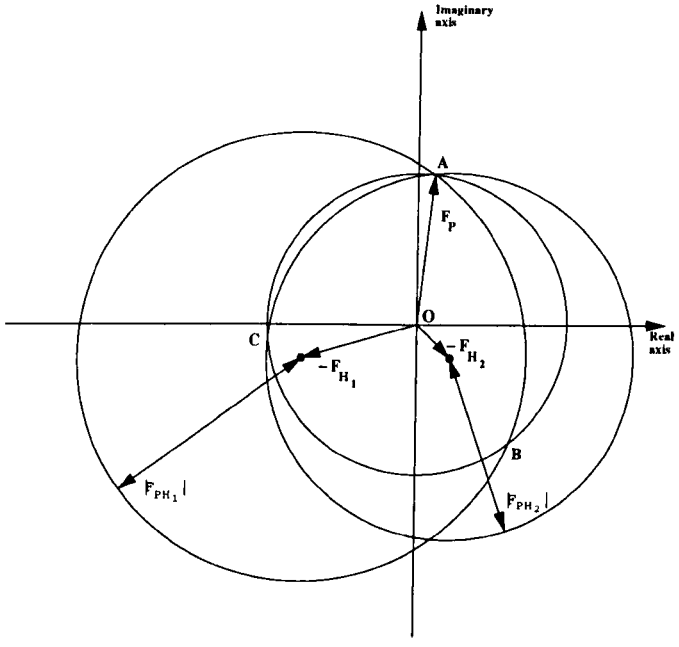

Figure 2.4: Phasing diagram for Multiple Isomorphous Replacement. See also legend to figure 2.3. When two derivatives are available, a third circle can be drawn, and in favourable circumstances the three circles intersect in one clearly defined point (point A in the diagram), thus resolving the phase ambiguity.

frequency of the radiation, and is of the form

$$f_j = (f_j)_C + \Delta f' + i\Delta f'' \tag{2.24}$$

$(f_j)_C$ is the centrosymmetrical approximation. $\Delta f'$ is the dispersion component and does not involve absorption. It is responsible for the change in refractive index away from unity. $\Delta f''$ is responsible for the absorption coefficient of the medium. This additional information causes the breakdown of Friedel's Law (see equation 2.15) and can be used to resolve the phase ambiguity, as can be seen in figure 2.5.

Without anomalous scattering, $|(F_{hkl})_{PH}| = |(F_{\bar{h}\bar{k}\bar{l}})_{PH}|$ (Friedel's Law) and $|(F_{\bar{h}\bar{k}\bar{l}})_{PH}| = |(F_{\bar{h}\bar{k}\bar{l}})^*_{PH}|$ (complex conjugates). In the presence of anomalously scat-

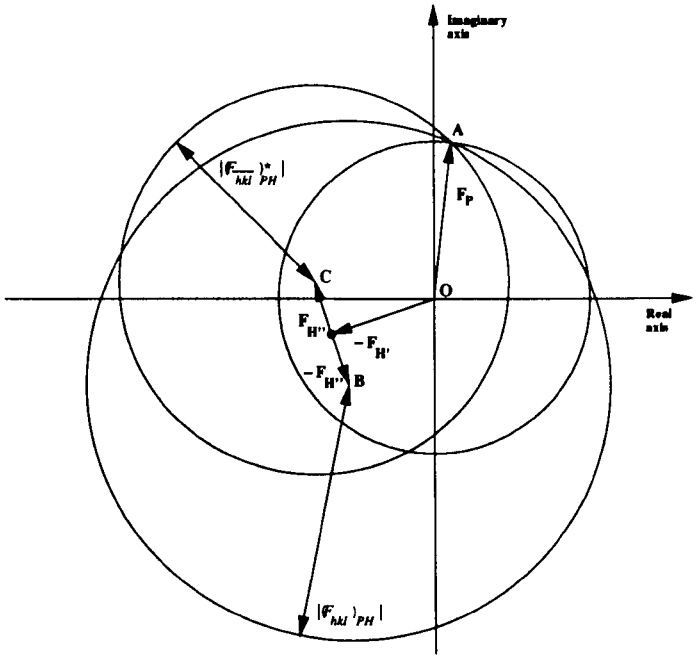

Figure 2.5: Phasing diagram for SIR with Anomalous Dispersion. See also legend to figure 2.3. The effect of anomalous scattering is to introduce a phase shift, which is different for reflections forming a Friedel pair. Thus there are circles with radii $|(F_{hkl})_{PH}|$ and $|(F_{\bar{h}\bar{k}\bar{l}})^*_{PH}|$ centred at $-(\mathbf{F}_H + \mathbf{F}''_H)$ and $-(\mathbf{F}_H - \mathbf{F}''_H)$ (points B and C in the diagram), intersecting with the circle corresponding to $|F_P|$ in only one point (point A in the diagram).

tering heavy atoms, the contribution of the native protein to $(\mathbf{F}_{hkl})_{PH}$ and $(\mathbf{F}_{\bar{h}\bar{k}\bar{l}})^*_{PH}$ is identical both in magnitude and phase. The magnitudes of the complex conjugates are still equal. The real component of $\mathbf{F}_H$, namely $\mathbf{F}'_H$ (based on $(f_j)_C + \Delta f'$ of the atomic scattering factor) is the same for $(\mathbf{F}_{hkl})_{PH}$ and $(\mathbf{F}_{\bar{h}\bar{k}\bar{l}})^*_{PH}$. The difference is in the imaginary component of $\mathbf{F}_H$, being $\mathbf{F}''_H$, which has different signs, so that $|(F_{hkl})_{PH}|$ and $|(F_{\bar{h}\bar{k}\bar{l}})^*_{PH}|$ sweep different circles. They must, however, intersect the $|F_P|$ circle in one point, which will represent the true value of $\mathbf{F}_P$.

In the case of insulin, anomalous scattering data was used in the form of what is called the 'anomalous Patterson', which uses $(|F_{hkl}| - |F_{\bar{h}\bar{k}\bar{l}}|)^2$ as coefficients. This

method was first used for proteins by Rossmann (Rossmann, 1961) and produced information of the distances between anomalously scattering atoms.

2.4.2 Molecular replacement

Molecular replacement methods were first developed by Rossmann and Blow (1962) to exploit the presence of non-crystallographic symmetry to obtain phase information and reduce phase uncertainties. The ideas were extended by Bricogne (1976) for real-space symmetry averaging to improve the electron density. The methods are based on a simple idea: finding a molecule in different places in a unit cell.

More recently, the term 'molecular replacement' is often used for the process of finding a similar molecule in a different crystal. If the structure of a molecule related to that under study is already known, it can be used to arrive at starting phases through orientation and positioning of the known molecule into the unit cell of the unknown. The basic principle of molecular replacement is rooted in the Patterson function. A Patterson map contains information of two types of vectors:

1. intramolecular or self-Patterson vectors of pairs of atoms in the same molecule. These vectors are relatively short and are thus clustered around the origin;

2. intermolecular or cross-Patterson vectors, which are generally longer than self vectors.

The self vector cluster would be equal for non-crystallographically related molecules in the same unit cell but also very similar for similar molecules in different crystals, apart from a rotation difference. The cross vectors provide information on the required translation of molecules to their positions relative to symmetry elements. Thus the process of molecular replacement can be divided into two: orientation and positioning, obliterating the need for a direct six-dimensional search, which is computationally unfeasible.

The rotation function

The construction of a Patterson function can be regarded as consecutively placing each atom in the origin and drawing vectors from such an atom to all other atoms in the unit cell. The value assigned to the vector is proportional to the product of the atomic numbers at either end of the vector. Thus, the height of the origin peak in a Patterson map is proportional to the sum of the squared atomic numbers, due to the null-vectors from each atom to itself. As stated before, the cross vectors in a protein Patterson are generally longer than the self vectors, unless the protein molecules are unusually closely packed or non-globular in shape.

In the self-rotation function, the Patterson of the unknown structure is rotated with respect to itself. Maximum overlap of these Patterson maps will occur at zero rotation (obviously) and at rotations representing (non-crystallographic) symmetry. The overlap is calculated by integrating the product over an appropriate spherical part of Patterson space (Rossmann and Blow, 1962). Similarly, two Patterson maps of related molecules (one known, one unknown) can be superposed when performing a cross-rotation function. The origin peak may be excluded by using an inner limit for the integration. The integration radius used should reflect the dimensions of the molecule and packing considerations, in order to exclude cross vectors. In the cross-rotation function the occurrence of cross vectors of the known molecule can be prevented altogether by putting the molecule in an artificially large P1 cell, so that cross vectors lie well outside the integration radius.

The general notation for the rotation function, seen as an overlap function of a Patterson function $P(\vec{u})$ and a rotated Patterson function $P_r(\vec{u}_r)$, dependent on Eulerian angles α, β, γ, is

$$R(\alpha, \beta, \gamma) = \int_U P(\vec{u}).P_r(\vec{u}_r)d\vec{u} \tag{2.25}$$

where U is the appropriate volume of integration, and $\vec{u}$ is a Patterson vector

(u, v, w).

If there is any prior knowledge of non-crystallographic symmetry, this can be used as a constraint in the calculation of what is named the 'locked' rotation function, a procedure first described by Rossmann *et al.* (1972). Assumed point-group symmetry can be imposed on the unknown molecule, improving the rotation function. The explicit use of non-crystallographic symmetry acts as a way of averaging the previously separate rotation functions R_i (Tong and Rossmann, 1990):

$$R_{LOCKED} = \frac{\sum_{i=1}^{n} R_i}{n} \tag{2.26}$$

where n is the number of non-crystallographic symmetry elements.

The rotation functions used in the studies of the structures in this thesis, were performed with the program package *AMoRe* (Navaza, 1994). The rotation function in *AMoRe* (Navaza, 1987) is based on the fast rotation function, using spherical harmonics and Bessel function expansions (Crowther, 1972), the detailed mathematics of which are beyond the scope of this thesis. The adaptations made by Navaza permit more accurate computation of the rotation matrices (Navaza, 1990), and enhance the resolution of the rotation function peaks. Removal of the origin (usually performed by using an inner limit for the integration radius) is achieved by excluding the lower order spherically symmetrical Bessel functions.

AMoRe uses an uncommon philosophy at the stage of the structure factor calculation from the model electron density. The model is rotated so that its principle axes of inertia are aligned with the cell axes and then placed, with its centre of mass, at the origin of a P1 cell with linear dimensions around four times the size of the model. Continuous Fourier coefficients are then calculated on a fine grid. Therefore the problem of the need for sampling at non-integral reciprocal lattice points during the calculation of the rotated Patterson function, does not occur. The structure factors can be interpolated from the continuous set, rather than being calculated

for every rotation. The loss of accuracy is small: only 1% in terms of crystallographic R-factor for the situation of a quadruple-size box. For a double-size box this discrepancy would be 8% (Castellano *et al.*, 1992).

The program outputs a list of rotation function peaks with Eulerian angles α, β, γ and a correlation coefficient. Because the model has been placed at the origin and rotated, the output is not in straightforward Euler angles. It needs to be shifted back, away from the origin, after which it may, in some cases, reveal directly interpretable information about the molecule under study.

The translation function

The first fast translation function expression was devised by Crowther and Blow (1967):

$$T_{jk}(\vec{v}) = \int_V P_{jk}(\vec{u}, \vec{v}).P_{obs}(\vec{u})d\vec{u} \tag{2.27}$$

$P_{jk}(\vec{u}, \vec{v})$ is the cross-Patterson function of the model structure in which two molecules (j and k) are related by crystallographic symmetry. $\vec{v}$ is the intermolecular vector between the local origins of these two molecules. $P_{obs}(\vec{u})$ is the Patterson function of the unknown structure. This function effectively searches one pair of asymmetric units, *i.e.* one crystallographic symmetry element, at a time. The function has two limitations:

1. The self vectors confuse the picture. They can easily be removed by subtraction, resulting in the T1 function:

$$T_{jk}1(\vec{v}) = \int_V [P_{obs}(\vec{u}) - \sum_i P_{ii}(\vec{u})].P_{jk}(\vec{u}, \vec{v})d\vec{u} \tag{2.28}$$

   assuming the intramolecular vector set of the unknown structure is the same as that of the model;

2. The solution of the translation function only represents one symmetry operator; all other symmetry information (including non-crystallographic symmetry) is lost. An improved translation function has been derived, resulting in peaks for all possible intermolecular vectors. The model now has as many molecules in the cell as the unknown structure. It is more convenient to express this translation function in terms of the translation vector $\vec{t}$ for the identity asymmetric unit. The expression contains the calculated Patterson for the correctly oriented model with unknown translation:

$$T0(\vec{t}) = \int_V P_{obs}(\vec{u}).P_{calc}(\vec{u},\vec{t})d\vec{u} \tag{2.29}$$

Both improvements can be incorporated together in the most commonly used T2 translation function:

$$T2(\vec{t}) = \int_V [P_{obs}(\vec{u}) - \sum_i P_{ii}(\vec{u})].[P_{calc}(\vec{u},\vec{t}) - \sum_i P_{ii}(\vec{u})]d\vec{u} \tag{2.30}$$

Computation of translation functions is usually performed in reciprocal space, sometimes using normalised structure factor amplitudes

$$\mathbf{E}_{hkl} = \frac{\mathbf{U}_{hkl}}{\langle |U_{hkl}|^2 \rangle^{1/2}} \tag{2.31}$$

where $\mathbf{U}_{hkl}$ is the unitary structure factor

$$\mathbf{U}_{hkl} = \frac{\mathbf{F}_{hkl}}{\sum_j f_j} \tag{2.32}$$

and $\langle |U_{hkl}|^2 \rangle$ is the average value of $|U_{hkl}|^2$. The normalised partial structure factor for the j'th asymmetric unit is related to the symmetry-equivalent structure factor for the identity by a simple phase shift, so there is no need to completely recalculate the structure factors every time.

Translation function calculations can be improved by

- including existing phase information from isomorphous replacement studies, calculating the so-called 'phased translation function' (Read and Schierbeek, 1988);
- including information of a partial model by either adding or subtracting partial Patterson functions appropriately during rotation or translation function calculations (Zhang and Matthews, 1994);
- improving the search model and its orientation before the translation search. Brünger (1990) has suggested the procedure of Patterson Correlation refinement, minimising a target function based on the correlation between observed and calculated Patterson functions and including an empirical energy term to restrain the geometry of the model. The model can be divided into domains treated as rigid bodies, while the domains are allowed to move with respect to each other;
- combining the translation search with a limited systematic variation of the parameters from the rotation function, assessing improvement by R-factor and correlation coefficient calculations (Fujinaga and Read, 1987);
- fixing one molecule in the asymmetric unit while searching for non-crystallographically related molecules (as in *AMoRe* (Navaza, 1994)).

The translation function may be by-passed altogether by a trial and error search, moving the correctly oriented model through the asymmetric unit and calculating the conventional R-factor as a function of the molecular position

$$R = \frac{\sum_{hkl} ||F_{obs}| - k|F_{calc}||}{\sum_{hkl} |F_{obs}|} \tag{2.33}$$

During these R-factor search calculations, the Fourier transform of the molecule only needs to be calculated once, for translations correspond to a pure phase shift which

can be applied directly. A slight improvement would be to calculate standard linear correlation coefficients, which are insensitive to structure factor scaling, rather than R-factors:

$$SLCC = \frac{\sum_{hkl}(|F_{obs}|^2 - < |F_{obs}|^2 >).(|F_{calc}|^2 - < |F_{calc}|^2 >)}{[\sum_{hkl}(|F_{obs}|^2 - < |F_{obs}|^2 >)^2 . \sum_{hkl}(|F_{calc}|^2 - < |F_{calc}|^2 >)^2]^{1/2}} \quad (2.34)$$

AMoRe uses straightforward T0 translation functions (Navaza and Vernoslova, 1995), with the option to incorporate information concerning already placed models (Navaza, 1994), enabling the positioning of additional molecules in an asymmetric unit. The program outputs α, β, γ, a translation vector (x, y, z) and an R-factor and correlation coefficient, the rotation and translation information again in "internal *AMoRe* format" (*vide supra*), which can be converted back to standard Euler angles.

More about *AMoRe*

A very powerful component of *AMoRe* is its fast rigid-body refinement procedure, which is very efficient in the evaluation of the correctness of the molecular replacement solution. The ideas on which it is based, were first developed by Huber and Schneider (1985). It is not a least-squares rigid-body refinement of coordinates, but rather a minimisation of the misfit, where rotations and translations are calculated by interpolation and phase shifts.

The sub-program (FITING) is described by Castellano *et al.* (1992). The minimisation procedure involves derivatives which are calculated from analytical expressions rather than numerical differentiation as used by Huber and Schneider (1985). The radius of convergence is around 10° for the angles and half the resolution of the data for the translations (Castellano *et al.*, 1992).

2.5 Refinement of protein structures

After determination of an initial model, either through molecular replacement or isomorphous replacement, the structure needs to be refined in order to obtain a set of atomic coordinates that corresponds best to the observed data. Various systematic and random errors have an effect on the accuracy of initial models. These errors have their origin for example

- during the data collection experiment. With the advances in data collection technology, these errors will become less significant;
- in the weak data (discussed in section 2.3);
- in the corrections during data processing (see section 2.3);
- while solving the Phase Problem. These errors are most prominent in the isomorphous replacement method, where 'lack of closure error', 'best phase' and 'phase probability' are terms frequently encountered (Blow and Crick, 1959, Terwilliger and Eisenberg, 1987, Read, 1991). The phases are inaccurate, so the map in which the initial model is built, is often not fully interpretable. In molecular replacement studies there are limitations to the efficiency of the functions used (Navaza, 1994). Moreover, in many cases the search model for the molecular replacement calculations is, at least locally, significantly different from the molecule under study, leading to a sub-optimal starting point for refinement;
- in determining certain absolute lengths (*e.g.* cell axes, bonds). Engh and Huber (1991) have determined a set of ideal bond lengths and angles *etc.* on the basis of small molecule structure data.

Therefore, the initial model is only an approximation, and the structure factors calculated from the model are in poor agreement with the observed structure factors.

This agreement is commonly stated in terms of the reliability index R:

$$R = \frac{\sum_{hkl} ||F_{obs}| - k|F_{calc}||}{\sum_{hkl} |F_{obs}|} \tag{2.35}$$

where k is the scale factor between $|F_{obs}|$ and $|F_{calc}|$. Wilson (1950) has shown that for an entirely incorrect non-centrosymmetric structure R=0.586. During the course of refinement, progress is followed by calculating the reliability index at various stages. This R-factor may reach artificially low values due to model bias introduced during the refinement procedure. From the statistical method of cross validation a new method for avoiding bias in macromolecular structure refinement has been developed by Brünger (1992). It is based on the idea that a fully refined model should be equally valid if new data is introduced. In the statistical cross validation method it is common to omit a single data point from the data, after which refinement takes place, rather than obtaining new data after refinement. Since the information from one data point is statistically unreliable, many data points have to be omitted in turn and the remaining information has to be fully refined. This full protocol would be impractical in general, because it would involve as many refinements as there are observations. In protein crystallography, however, the data usually exhibit a certain degree of redundancy, which means that a statistically relevant number of data may be omitted from the data prior to refinement. These data can then be used to follow the progress of refinement by means of the 'free R-factor', as introduced by Brünger (1992):

$$R_T^{free} = \frac{\sum_{hkl \in T} ||F_{obs}| - k|F_{calc}||}{\sum_{hkl \in T} |F_{obs}|} \tag{2.36}$$

R_T^{free} refers to the 'test set T', a set of reflections (commonly 5-10% of the observed data) which are excluded from refinement. Refinement is carried out with the remaining reflections only, called 'working set W'.

Refinement is usually performed by 'least-squares' minimisation, *i.e.* the sum of the squared differences between calculated and observed structure factors is min-

imised. In traditional least-squares methods in protein crystallography, the function to be minimised is commonly chosen as

$$Q = \sum_{hkl} w(hkl)(|F_{obs}| - k|F_{calc}|)^2 \tag{2.37}$$

Q is called the crystallographic residual. The summation is taken over all independent reflections hkl, and $w(hkl)$ is the weight applied to each individual observation. $w(hkl)$ can be any weighting scheme, but is usually, at least initially, chosen to be 1. Some computing packages have the option of use of σ, the standard deviation of the observation hkl, in the form of $\frac{1}{\sigma^2(hkl)}$. The minimum of Q is reached when the first derivative with respect to each parameter p_j is zero:

$$\frac{\partial Q}{\partial p_j} = 0 \tag{2.38}$$

or

$$\sum_{hkl} w(hkl)\Delta\frac{\partial|F_{calc}|}{\partial p_j} = 0 \tag{2.39}$$

where $\Delta = |F_{obs}| - |F_{calc}|$ (assuming absolute scale). Starting from a set of parameters $\mathbf{p}$ and necessary corrections ξ, Δ can be expanded as a Taylor series to the first order:

$$\Delta(\mathbf{p} + \xi) = \Delta(\mathbf{p}) - \sum_{i=1}^{n} \xi_i \frac{\partial|F_{calc}|}{\partial p_j} \tag{2.40}$$

for n parameters. Combining equations 2.39 and 2.40 results in normal equations of the form:

$$\sum_{i=1}^{n}\left[\sum_{hkl} w(hkl)\frac{\partial|F_{calc}|}{\partial p_i}\frac{\partial|F_{calc}|}{\partial p_j}\right]\xi_i = \sum_{hkl} w(hkl)\Delta\frac{\partial|F_{calc}|}{\partial p_j} \tag{2.41}$$

The n normal equations may be expressed in matrix form:

$$\mathbf{A}\xi = \mathbf{b} \quad or \quad \sum_i a_{ij}\xi_i = b_j \tag{2.42}$$

where

$$a_{ij} = \sum_{hkl} w(hkl) \frac{\partial |F_{calc}|}{\partial p_i} \frac{\partial |F_{calc}|}{\partial p_j} \tag{2.43}$$

and

$$b_j = \sum_{hkl} w(hkl) \Delta \frac{\partial |F_{calc}|}{\partial p_j} \tag{2.44}$$

These equations are solved by matrix inversion for the set of shifts ξ which are then applied to the estimates $\mathbf{p}$. The process is repeated until convergence is reached, *i.e.* when the ξ_i become negligible.

In protein crystallography the sheer size of matrix $\mathbf{A}$ prohibits calculation of all the elements and full inversion. A method must be chosen to approximate these steps. In the various refinement packages available today, different ways have been chosen for the minimisation:

- Simulated annealing, used by X-PLOR (Brünger *et al.*, 1987) in the early stages of refinement, while later on it uses
- Conjugate-gradient, a particular variant of the conjugate-direction method (Fletcher and Reeves, 1964). This is used by TNT (Tronrud *et al.*, 1987) and REFMAC (Murshudov *et al.*, 1996a). TNT and REFMAC also have the option of using the more general conjugate-direction method;
- Diagonal approximation to the normal matrix, used by SFRF (Agarwal, 1978) and EREF (Jack and Levitt, 1978);
- Sparse-matrix approximation. The sparse matrix may then be inverted, which in PROLSQ (Hendrickson and Konnert, 1980, Konnert and Hendrickson, 1980, Hendrickson, 1985), REFMAC and CORELS (Sussman *et al.*, 1977, Sussman, 1985) is done by conjugate-gradient methods.

The success of protein structure refinement depends largely on the degree of overdetermination of the system, *i.e.* the number of observed data for every parameter to be refined. The ratio of observed data over parameters may be changed

favourably either by providing additional data (often the incorporation of stereochemical knowledge in the form of restraints) or by reducing the number of parameters (often by constraining features of the structure to specific values). In the case of restrained refinement, in practice the function minimised is then a composite of diffraction and stereochemical information:

$$Q_{composite} = Q + Q_{stereo} \tag{2.45}$$

In PROLSQ Q_{stereo} contains terms for bonding distances, planar groups, chiral centres, nonbonded contacts and torsion angles. The formalism of equation 2.45 can also be used for the description of the minimisation function in refinement packages other than PROLSQ. For example, in CORELS Q would be the usual sum of structure factor differences, while Q_{stereo} would be replaced by three terms, namely for all interatomic distances, nonbonded close contacts and a term to prevent the structure from moving away too far from the input coordinates (Sussman, 1985). In EREF again the Q represents the crystallographic term, and Q_{stereo} takes on the form of a conformational energy term consisting of contributions for bonds, bond angles, torsion angles and nonbonded interactions as described by Levitt (1974). Each term in $Q_{composite}$ requires appropriate weighting and scaling.

In general, least-squares refinement as described above, has a limited radius of convergence. It generally does not correct errors in atomic coordinates which are larger than 1Å and is easily trapped in local minima. To avoid this, a refinement technique is required which will allow up-hill steps away from the nearest minimum. Simulated annealing is such a method. Metropolis and co-workers (1953) introduced an algorithm to simulate a collection of atoms at a given temperature. In each step of the simulation an atom is moved at random and the effect on the total energy is computed. If the energy decreases, the move is accepted, if it increases, it is accepted with a probability which is proportional to $exp[-\Delta E/k_bT]$ (ΔE is the energy differ-

ence between the two states, k_b is the Boltzmann constant and T the temperature). When the energy function is replaced by the cost function of a minimisation problem, and the atomic parameters by the parameter set ξ, the Metropolis procedure can be used to generate a set of configurations at a given temperature of the minimisation problem. The temperature parameter determines the amount of energy available to the system to overcome energy barriers, *i.e.* at a higher temperature more steps are accepted which lead to an increase in the cost function. This allows the minimisation to escape from local minima. Simulated annealing is a process in which the temperature is first raised to a relatively high value, after which the system is slowly cooled until it freezes and no more changes occur. The maximum temperature, the rate of cooling, and the number of rearrangements of ξ determine the 'annealing schedule'. A direct application of the Metropolis algorithm to the refinement of macromolecules is very inefficient: most of the random changes in ξ would lead to violation of covalent bonds. Instead, molecular dynamics is used to follow the gradients of a target function which includes an empirical energy term to introduce stereochemical and other restraints. Thus the total energy is the sum of an empirical energy describing the restraints and an effective potential energy to include the X-ray information. The empirical energy contains terms for bond lengths, bond angles, dihedral torsion angles, chiral centres, planarity of aromatic rings and Van der Waals and electrostatic interactions:

$$E_i = E_{bonds} + E_{angles} + E_{dihedrals} + E_{chiral,planar} + E_{vdWaals,elec} \tag{2.46}$$

The program package X-PLOR essentially uses the stereochemical and nonbonded parameters as formulated for the molecular dynamics program CHARMM (Brooks *et al.*, 1983). Conventional refinement is performed using a conjugate-gradient minimisation algorithm. The application of simulated annealing consists of 'heating and cooling' the structure. An initial temperature, typically 3000K, is applied to

the system. Two protocols are possible: either the system is extensively equilibrated at this temperature after which cooling takes place, or the system is allowed to cool directly, very slowly. In the first protocol the initial velocities are assigned from a Maxwellian distribution at the appropriate temperature. The temperature is controlled through rescaling of the velocities at set intervals of the simulation. The scale factor is determined from:

$$Wa = \frac{\langle \sum_i m V_i^{old}(t)^2 \rangle}{n k_b T} \tag{2.47}$$

where m is the mass and V_i^{old} the velocity of atom i at time t, k_b is the Boltzmann constant, T the temperature, n the number of degrees of freedom and $\langle\rangle$ the average over the time interval between rescaling of the velocities. It has been shown (Weis and Brünger, 1989) that towards the end of refinement a reduction of the weight by 10-40% is beneficial, in order to prevent deterioration of the model geometry. Cooling of the system is effected through stepwise lowering of the temperature. At each temperature the system is allowed to equilibrate employing the rescaling as expressed in equation 2.47.

A large number of applications (for an extensive overview see Goodfellow *et al.*, 1989) has shown that the shifts observed are very much larger than for conventional refinement methods. This leads to dramatic improvements for initial models of macromolecular structures. Although the radius of convergence of refinement with simulated annealing/molecular dynamics techniques is much increased, manual intervention in the form of map inspection and model rebuilding is still necessary.

Until a few years ago, manual rebuilding was mainly performed with the graphics program FRODO (Jones, 1978). Recently, the molecular modelling package Quanta (Molecular Simulations, 1996) has been extended with a large crystallographic module, containing the application X-AUTOFIT (Oldfield, 1996). X-AUTOFIT is designed to enhance the efficiency of *de novo* map building as well

as general model building in later stages of macromolecular refinement. The model building is carried out with the aid of real space refinement, regularisation and rigid body refinement algorithms, as well as traditional manual editing facilities. Three refinement techniques are implemented: grid searching about torsions, torsion angle real space gradient refinement and Monte Carlo fitting. For the automatic building and refining of ligand and water coordinates Quanta provides the applications X-LIGAND and X-SOLVATE, respectively. X-LIGAND is designed to search for unsatisfied electron density, sort these in order of volume and fit a ligand automatically. It can be used to fit small (in)organic molecules (such as sulfate ions and phenol molecules), polysaccharides and small polypeptides. X-SOLVATE allows the search of an electron density map for water peaks in a specified space around the protein.

Building up the water structure is also very conveniently performed with the program ARP (Lamzin and Wilson, 1992). The program seeks to improve the fit of atomic coordinates to electron density maps by appropriate removal and addition of atoms. This procedure is alternated with reciprocal space refinement of the complete set of coordinates with, for example, PROLSQ or REFMAC.

2.6 Maximum entropy and maximum likelihood

2.6.1 The concept of entropy

From information theory came the realisation that there is a unique, unambiguous criterion for the amount of uncertainty represented by a probability distribution $\mathbf{q}(\mathbf{x}) = \sum_{i=1}^{N} x_i$ (N is the number of states of x_i, and $\sum_i q_i = 1$) or $\int_V x dx$ for a continuous distribution in volume V. This quantity

$$S(q_i) = -\sum_{i=1}^{N} q_i ln q_i \tag{2.48}$$

is continuous in q_i, increases with increasing uncertainty and is additive for independent sources of uncertainty (Shannon and Weaver, 1949). In the simple case where the system consists of two events x and y, it can be shown that

$$S(x,y) \leq S(x) + S(y) \tag{2.49}$$

Or, in words, the uncertainty of a joint event is less than or equal to the sum of the individual uncertainties (equality only occurring if the events are independent). Also, the probability of event y is influenced by the occurrence of event x, resulting in a conditional entropy for event y, $S_x(y)$. It then follows that

$$S(x,y) = S(x) + S_x(y) \tag{2.50}$$

Thus the uncertainty of the joint event x, y is the uncertainty of x plus the uncertainty of y when x is known. Combining equations 2.49 and 2.50, it can be seen that $S(y) \geq S_x(y)$, or, the uncertainty of y is never increased by knowledge of x. It will be decreased unless x and y are independent events, in which case it is not changed.

The quantity S is equivalent to the expression for entropy as found in statistical mechanics (Jaynes, 1957). The minimum value of the entropy, $S = 0$, is reached when one of the probabilities q_i is 1, *i.e.* when there is no uncertainty. In that case there is no freedom of choice and no information is gained from the situation. The unconditional maximum value for S, namely lnN, is reached when all parameters x_i are equally probable, *i.e.* for a uniform distribution. Thus S increases when the q_i are more 'average'.

2.6.2 Maximum entropy

Whenever an experiment is undertaken, there is some form of prior knowledge, even if it is only intuitive. Jaynes (1979) writes: "merely knowing the physical meaning of [the] parameters [of our experiment] already constitutes highly relevant prior information which our intuition is able to use at once. Can we analyse how our intuition does this, extract the essence, and express it as a formal mathematical principle that might apply in cases where our intuition fails us?". In fact, an experiment in itself constitutes information, which has an effect on the entropy of the system under study.

In most experiments, the unconditional maximum value for S, as described in the previous section, will not be reached, simply because of the requirements the prior knowledge puts on the experiment. If testable information is available, the process of finding the maximum value of S will be put under constraints. Here 'testable information' is defined as: such that it may be determined unambiguously whether the information does or does not agree with the proposed tested distribution p_i, the posterior distribution. An example of testable information is: "It is certain that $\cos\alpha < 0.7$". Untestable information (*e.g.* "The mean value of $\cos\alpha$ might be less than 0.7") can at present not be used in mathematical theory. Jaynes, however, believes that new principles for this must exist (Jaynes, 1979).

In an experiment undertaken to test information, linear constraints may be given in the following form:

$$\sum_{i=1}^{N} p_i f_k(x_i) = F_k \qquad for\ k = 1, m \tag{2.51}$$

where $\sum_i p_i = 1$, which is in fact the zero-order constraint. F_k is the expected value of f_k in the experiment (see example I). $f_k(x_i)$ is a set of m functions of the parameters x_i. If $m < N$, the maximum entropy may be calculated. The derivation of the formula for the maximum entropy may be done as follows (for the

mathematical principle of 'Lagrange multipliers' see appendix A):

$$p_i = \frac{1}{Z(\lambda_1 ... \lambda_m)} exp[\lambda_1 f_1(x_i) + ... + \lambda_m f_m(x_i)] \tag{2.52}$$

where

$$Z(\lambda_1 ... \lambda_m) \equiv \sum_{i=1}^{N} exp[\lambda_1 f_1(x_i) + ... + \lambda_m f_m(x_i)] \tag{2.53}$$

is the partition function, which is a normalisation function so that the sum of the p_i is 1. λ_k are the Lagrange multipliers, which are chosen so as to satisfy the constraints 2.51, namely

$$F_k = \frac{\partial}{\partial \lambda_k} lnZ \qquad for\ k = 1, m \tag{2.54}$$

This is a set of m equations for m unknowns. The resulting maximum entropy is a function of the prior information:

$$S(F_k) = lnZ - \sum_k \lambda_k F_k \tag{2.55}$$

Example I

An example of the use of entropy in sciences other than physics was given by Jaynes in 1962, and was reproduced in an extensive article called "Where do we stand on Maximum Entropy?" (Jaynes, 1979). A seemingly ordinary die has been tested prior to a real experiment, and the information from this test is: the average number of spots up is not 3.5 as in an 'honest' die, but 4.5. What is the probability assigned to i spots ($1 \leq i \leq 6$) on the next throw? The data may be interpreted in the form of a constraint:

$$\sum_{i=1}^{6} i p_i = 4.5 \tag{2.56}$$

The partition function is $Z(\lambda) = \sum_i e^{\lambda i} = x(1-x)^{-1}(1-x^6)$ with $x \equiv e^{\lambda}$. Equation 2.56 becomes $\frac{\partial}{\partial \lambda} lnZ = \frac{1-7x^6+6x^7}{(1-x)(1-x^6)} = 4.5$ or $3x^7 - 5x^6 + 9x - 7 = 0$. The desired root is x=1.44925, giving λ=0.37105 and Z=26.66365. The maximum entropy probabilities

are $p_i = Z^{-1}x^i$, or $\{p_1...p_6\}$=$\{0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749\}$. This is the distribution based on the least assumptions, *i.e.* the distribution which is nearest to being uniform under the constraints presented in the description of the problem. The entropy of this distribution is S=1.61358. The unconditional maximum entropy, *i.e.* the entropy of an honest die, is $ln6$=1.79176, representing a system with no constraints and a totally uniform distribution.

2.6.3 Relative entropy

The availability of prior knowledge implies the existence of a prior probability distribution $\mathbf{q}(\mathbf{x})$, which can be either discrete or continuous. After performing an experiment, a posterior probability distribution $\mathbf{p}(\mathbf{x})$ may be calculated. The relative entropy of the posterior distribution (the distribution of the outcome of the experiment) with respect to the prior distribution (*i.e.* before the experiment) is defined as:

$$S(\mathbf{p},\mathbf{q}) = -\sum_{i=1}^{N} p_i ln(p_i/q_i) \tag{2.57}$$

where N is the number of possible states of the system. This illustrates that the probability distribution over the system states is modified by the experiment, from prior distribution $\mathbf{q}(\mathbf{x})$ to posterior distribution $\mathbf{p}(\mathbf{x})$ (see example II).

Example II

An experimenter is given a set of dice, commonly expected to have q_i=1/6. An experiment of throwing the dice and calculating the entropy of each system is undertaken (see figure 2.6). From this example it can be seen that the more predictable the outcome is, the lower is the relative entropy of the system under investigation. Again totally predictable is seen as gaining no information (compare section 2.6.1).

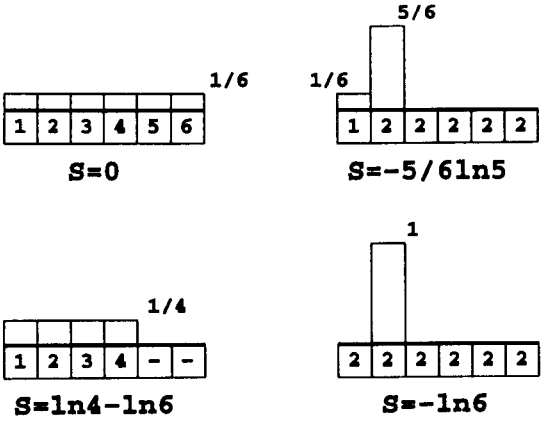

Figure 2.6: Throwing dice. Different types of dice and their associated distributions, indicating the correlation between entropy and information. 0ln0 is assumed to be 0. Example reproduced from Kinneging (1986).

The minimum entropy principle

The main goal of a successful experiment, is obtaining as much undoubtable data (to avoid confusion, the term 'information' has not been used here) as possible, leaving the least uncertainty. Entropy may therefore be used as a guideline in the choice of experiment. An experiment which provides maximum data has minimum entropy relative to any other experiment pertaining to the same question. These undoubtable minimum-entropy data provide a boundary for the entropy maximisation to obtain the best possible distribution based solely on trusted facts and no unprovable assumptions. In cases where the minimum-entropy experiment can not be chosen, the resulting data will not be optimal.

2.6.4 Bayes' theorem and likelihood

The use of prior knowledge and the inference of information from the observations of an experiment may also be dealt with entirely through statistics. To this end,

Bayes proposed the following theorem:

$$P(B|A) = P(B)P(A|B) \tag{2.58}$$

$P(B)$ is the *prior* probability of event B. $P(A|B)$ is the *conditional* probability of obtaining experimental information A in an experiment to test B. $P(A|B)$ modifies prior probability $P(B)$ in such a way that *posterior* probability $P(B|A)$ is obtained, the probability of B when A has been seen. This simple description implies that experimental information A and the connected probabilities constitute a learning experience about B. The right-hand side of equation 2.58 can be normalised for a set of events B_r, resulting in

$$P(B_r|A) = \frac{P(B_r)P(A|B_r)}{\sum_r P(B_r)P(A|B_r)} \tag{2.59}$$

In the next step of the learning process, the previous information A is seen as prior information H, after which new experimental information A will be obtained through a new experiment. The following form of Bayes' theorem is used at this stage:

$$P(B_r|A,H) = \frac{P(B_r|H)P(A|B_r,H)}{\sum_r P(B_r|H)P(A|B_r,H)} = \frac{P(B_r,A|H)}{\sum_r P(B_r,A|H)} \tag{2.60}$$

The terms of this equation may be explained as follows:

B_r is a set of mutually exclusive and exhaustive events (in the statistical sense), *i.e.* the knowledge tested through an experiment.

$P(B_r|H)$ is a *prior* probability, which is the probability of obtaining the parameters B_r solely on the basis of prior information H.

$P(B_r|A,H)$ is a *posterior* probability, *i.e.* the probability of obtaining new parameters B_r on the basis of prior information H and experimental information A.

$\frac{P(A|B_r,H)}{\sum_r P(B_r|H)P(A|B_r,H)}$ may be called the *likelihood* $\mathrm{L}(B_r)$. This is the probability that experimental information A is obtained in an experiment to study parameters B_r

when prior information H is available.

$P(B_r, A|H)$ is called the *joint* probability of B_r and A *conditional* on H.

Thus the theorem states that the probability that B_r occurs given the occurrence of A and currently available information H, is proportional to the probability of B_r on H multiplied by the probability of A on B_r and H. Or in other words, it gives the probabilities of B_r when A is known to have occurred.

Example III

The use of Bayes' Theorem may be explained by means of an example taken from Kendall's Advanced Theory of Statistics (1987). A box contains four dice, which are known to be either (a) all white, or (b) two white and two black. A die is drawn at random and found to be white. What is the probability that all the dice are white? Based on (a) and (b) there are two hypotheses B_1 and B_2. On B_1 the probability $P(A|B_r, H)$ (see equation 2.60) of getting a white die is 1, on B_2 it is $\frac{1}{2}$. From equation 2.60 it can be seen that

$$\begin{aligned} P(B_1|A,H) &= \frac{P(B_1|H)}{P(B_1|H)+\frac{1}{2}P(B_2|H)} \\ P(B_2|A,H) &= \frac{\frac{1}{2}P(B_2|H)}{P(B_1|H)+\frac{1}{2}P(B_2|H)} \end{aligned} \tag{2.61}$$

Bayes postulated that it is reasonable to take prior probabilities to be equal when nothing is known to the contrary, so $P(B_1|H) = P(B_2|H) = \frac{1}{2}$. Thus

$$P(B_1|A,H) = \frac{2}{3} \; and \; P(B_2|A,H) = \frac{1}{3} \tag{2.62}$$

The experiment continues by replacing the die and again drawing one at random. If it is found to be black, hypothesis (a) is rejected outright. If it is white, new posterior probabilities can be calculated in which the probabilities from equation

2.62 become priors: $P(B_1|H) = \frac{2}{3}$ and $P(B_2|H) = \frac{1}{3}$, where H includes A, the occurrences from the previous experiment. The new posterior probabilities based on the new event A' are

$$P(B_1|A', H) = \frac{4}{5} \; and \; P(B_2|A', H) = \frac{1}{5} \tag{2.63}$$

On repeating the process and again drawing a white die, the new posterior probability of (a) will be still higher, in agreement with common sense: repeated sampling (with replacement) without producing a black die increases the probability that there are no black dice present.

2.6.5 Entropy and likelihood in protein crystallography

In 1982 Collins proposed the use of iterative entropy maximisation for macromolecular crystallography. The ideas involved have their foundations in information theory and image enhancement, and their application has the potential to produce electron density maps of higher interpretability (through higher signal to noise ratio) than expected for Fourier synthesis of the available structure factors (Collins, 1982). Applying the general statements from sections 2.6.1 to 2.6.3 to crystallography it can be seen that

- in order to be able to obtain most information from a data collection experiment, the least uncertainties should be left (the minimum entropy principle) by collecting as much data as possible to the highest multiplicity and resolution that may reasonably be obtained;

- with the data available, the least assumptions should be made, so that the maximum entropy principle may be applied in the determination of the atomic structure, thus avoiding overfitting and bias.

Although the Phase Problem will not be solved experimentally in the foreseeable future, and is mathematically indeterminate when only amplitudes are known, some combinations of phases are more acceptable than others when solving a crystal structure. This is because the phases will have to yield a chemically valid distribution of atoms.

Bricogne (1984) proposed the use of a phasing method based on entropy maximisation and likelihood ranking as a means of applying probabilistic direct methods to macromolecules. The mere assumption that a crystal structure is made up of atoms, is enough to provide statistical restraints on the structure factors derived from the data, even if the atoms are randomly placed and independent of each other.

There are a number of advantages to maximum entropy methods (hereafter called MEM) over conventional direct methods (CDM). In principle the starting point of CDM and entropy/likelihood methods is the same, but the latter do not require a uniform distribution of the probability distribution of the atoms in the asymmetric unit (Bricogne and Gilmore, 1990). Also, MEM do not require high overdetermination by the data, unlike CDM (Bricogne, 1991a). Thirdly, CDM are most accurate for small **F**s because of the method of approximation of the probability distribution $P(\mathbf{F})$ (the joint probability distribution of all **F**s), whereas through MEM it is possible to use the more informative large **F**s (see below). Fourthly, in deriving $P(\mathbf{F})$ from the random atom model, CDM have, until recently, never made use of the complete $P(\mathbf{F})$ but instead considered individual terms for small groups of reflections which were combined (joined) later. However, Peschar and Schenk (1987) have addressed this problem for CDM.

Bricogne's Multisolution Tree

For a uniform prior distribution of atoms ($q(\vec{r})\equiv 1/V$) the joint probability distribution $P(\mathbf{F})$ may be calculated to various degrees of approximation. Observed magnitudes $|F|$ may be substituted into $P(\mathbf{F})$ after which only phases remain as free

variables, which yields a conditional distribution for the phases under the constraints of the observed $|F|$s. This conditional distribution should imply that certain combinations of phases are more probable than others. For small $|F|$s the conditional distributions of the phases will be least informative because they do not significantly narrow down the uncertainty in $\mathbf{F}$. With MEM 'recentering' is possible by setting up local approximations to $P(\mathbf{F})$ away from $|F| = 0$ by means of 'trial phase' shifts. This is equivalent to using a non-uniform prior distribution of atoms, the maximum entropy distribution $q^{ME}(\vec{r})$. From this description of recentering to obtain local approximations of $P(\mathbf{F})$ it may be seen it is not so much the maximum entropy distribution $q^{ME}(\vec{r})$ which is the really useful result, but the optimal approximation to $P(\mathbf{F})$. The maximum entropy distribution is almost just a by-product of the calculations of $P(\mathbf{F})$. The resulting joint probability distribution for all $\mathbf{F}$s is called $P^{SP}(\mathbf{F})$, the saddlepoint approximation to $P(\mathbf{F})$. In practice, the diagonal approximation to $P^{SP}(\mathbf{F})$ is used.

The first set of phases H chosen on the basis of the first conditional distribution may be seen as the origin-fixing phases and are assigned to the 'root node' of a 'solution tree', which will help to keep track of the phase choices made at various stages. At this node with corresponding basis set H the unique distribution q^{ME}, having maximum entropy compatible with the data attached to the node, is constructed by maximising the relative entropy

$$S_m(q) = -\int_V q(\vec{r}) ln \frac{q(\vec{r})}{m(\vec{r})} d\vec{r} \tag{2.64}$$

under the constraints

$$\int_V q(\vec{r}) exp(2\pi i \vec{h} \cdot \vec{r}) d\vec{r} = |U_{\vec{h}}|^{obs} exp[i\phi(\vec{h})] \qquad for\ all\ \vec{h} \in H \tag{2.65}$$

$m(\vec{r})$ is the normalised prior distribution of $q(\vec{r})$ which comes about because of the use of unitary structure factors $\mathbf{U}_{\vec{h}}$. Bricogne has named $m(\vec{r})$ the 'prior prejudice'

(Bricogne, 1984) and the 'probability density' (Bricogne and Gilmore, 1990). $\phi(\vec{h})$ is the phase of reflection $\vec{h}$.

Besides reproducing the amplitudes and phases attached to the node for reflections in H, q^{ME} has Fourier coefficients $\mathbf{U}_{\vec{k}}^{ME}$ with non-negligible amplitude for many non-basis reflections $\vec{k}$. This means that the conditional distribution of $|U_{\vec{k}}|$ deviates from the normal Gaussian distribution of Wilson statistics, which would correspond to $|U_{\vec{k}}^{ME}| = 0$, and that the atoms are distributed according to q^{ME} rather than uniformly. The two distributions just derived (q_0 and q^{ME}) under two different hypotheses on $\mathbf{U}_{\vec{h}}$ can not be easily compared because they involve phases $\phi(\vec{k})$ of non-basis reflections. However these phases may be integrated out to obtain the conditional probability distributions not of structure factors $\mathbf{U}_{\vec{k}}$ for reflections $\vec{k}$ in non-basis set K but of their moduli $|U_{\vec{k}}|$. Then the conditional probabilities assigned to the observed values of the moduli under both hypotheses may be compared, the better assumption being that which assigns the highest probability to them. This quantity acts as a figure of merit and is called the 'likelihood' of the hypothesis, defined as 'the probability it assigned to the actual result of an observation before that observation was performed' (Bricogne, 1991a). The comparison is best carried out using the 'likelihood ratio':

$$\frac{\mathrm{L}(\mathbf{U}_{\vec{h}}^{ME})}{\mathrm{L}(0)} = \frac{P(U_{\vec{k}}^{obs}|\mathbf{U}_{\vec{h}} = \mathbf{U}_{\vec{h}}^{ME})}{P(U_{\vec{k}}^{obs}|\mathbf{U}_{\vec{h}} = 0)} \tag{2.66}$$

where $U_{\vec{k}}^{obs} = |U_{\vec{k}}^{obs}|$. The likelihood ratio measures the extent to which the observed values of yet unphased moduli $|U_{\vec{k}}^{obs}|$ are made more probable by the assumption that $\mathbf{U}_{\vec{h}} = \mathbf{U}_{\vec{h}}^{ME}$ than by the assumption that $\mathbf{U}_{\vec{h}} = 0$. The logarithm of the likelihood ratio is called the 'log-likelihood gain', which is also often used.

Invoking Bayes' theorem the posterior probability distribution of $\mathbf{U}_{\vec{h}}^{ME}$ may be calculated by

$$P^{post}(\mathbf{U}_{\vec{h}}^{ME}) \propto P^{SP}(\mathbf{U}_{\vec{h}}^{ME}) \times \mathrm{L}(\mathbf{U}_{\vec{h}}^{ME}) \tag{2.67}$$

Then the basis-set phases in $\mathbf{U}_{\vec{h}}$ can be refined by maximising the posterior probability with respect to the phases.

From the descriptions so far, it can be summarised that the growth of the 'multisolution tree' has entailed:

1. choice of basis set H of origin-fixing (and/or enantiomorph-defining) reflections and assignment of the root node;

2. update of the prior distribution of atoms $q(\vec{r})$ to the maximum entropy distribution $q^{ME}(\vec{r})$ compatible with the phase choices, giving rise to structure factor values $\mathbf{U}_{\vec{h}}$ for that node;

3. construction of the conditional distribution P_{diag}^{SP} of the yet unphased structure factors in non-basis set K;

4. construction of the likelihood function L by integrating the conditional distribution over the unknown phases and substituting the observed values of the moduli in K;

5. refining the basis set phases in $\mathbf{U}_{\vec{h}}^{ME}$ by maximising L with respect to these basis phases.

The subsequent growth of the tree involves 'branching out' and 'pruning':

6. identification of the local maxima of the conditional distribution with respect to a subset of unknown phases when the values of the corresponding moduli are introduced;

7. expansion of the current node by creating a branch leading to a new tip node for each of these maxima;

8. in order to avoid development of spurious branches while maximising the chance of finding the correct set of phases, two criteria are used at each node:

- the loss of entropy in going from a uniform distribution of atoms $m(\vec{r}) = \frac{1}{V}$ to the current non-uniform $q^{ME}(\vec{r})$, which measures the shrinkage of the population of 'reasonably probable structures' when the constraints $\mathbf{U}_{\vec{h}}^{ME}$ are enforced (Shannon and Weaver, 1949);
- the likelihood, which measures the chance of survival of the node by measuring the degree to which the phase choices made anticipate correctly the information present in the yet inactive constraints.

The validity of the phase choices for H and some for K firmly established, the tree may be further developed from each surviving node by iteration from step 2. The entropy and the likelihood are complementary measures of the worth of the trial phase set, in that the entropy 'looks back' at the cost of accommodating the current phase assumptions, and the likelihood 'looks ahead' at the match between its predictions and the actual observations for yet unphased data.

Over the years Bricogne's theories have been updated with the incorporation of 'all sources of phase information normally used in macromolecular crystallography' (Bricogne, 1993) and the possibility to use the system for data other than from single crystal X-ray crystallography (*e.g.* powder diffraction data (Gilmore *et al.*, 1991)). Bricogne envisages the complete implementation of the 'Bayesian programme', *i.e.* the use of Bayesian theory in every stage of macromolecular structure determination, from first phasing (ultimately *ab initio*) to full refinement (Bricogne, 1993). Three program packages have been developed as a direct result of these theories, namely MICE (Bricogne and Gilmore, 1990), BUSTER (Bricogne, 1991b, Bricogne, 1991c) and SHARP (de la Fortelle and Bricogne, 1996).

REFMAC

More recently, the program REFMAC (Murshudov *et al.*, 1996b) has been developed, to perform maximum likelihood refinement. The likelihood function (*vide*

infra) is constructed from available parameters at any point in a refinement procedure, whether immediately after structure determination or much later on in refinement. Instead of maximisation of the likelihood function, an equivalent calculation is performed: the logarithm of this function (the likelihood residual) is minimised to a desired degree of convergence through conventional methods. This procedure of constructing the likelihood function and then minimisation may be seen as a macrocycle of likelihood refinement, which may be iterated to convergence itself. The program uses prior knowledge of the stereochemistry of macromolecules, like in least-squares methods, but works with conditional probability distributions rather than the Gaussian distributions employed in least-squares. Partial phase information may be added and there is an option to use the information provided by experimental standard deviations. Different parts of a structure may be treated in different ways if the errors in the separate domains are known or suspected to be different. Bulk solvent scaling according to Tronrud (1996) has been incorporated.

When experimental data (*e.g.* observed magnitudes $|F|$) are available and parameters x (*e.g.* atomic coordinates) are to be estimated, Bayes' theorem may be written as:

$$P(x;|F|) = q(x)\frac{P(|F|;x)}{P(|F|)} = q(x)\mathrm{L}(x;|F|) \tag{2.68}$$

where ';' denotes conditionality to avoid confusion with the '|' in $|F|$. $P(x;|F|)$ is the posterior probability of x on the basis of the data $|F|$; $q(x)$ is the prior distribution of the atoms. In theory, this could be the maximum entropy distribution q^{ME} as described in the previous section. However, REFMAC does not deal with maximum entropy calculations. Therefore, $q(x)$ is typically the distribution of a Molecular or Isomorphous Replacement model.

$$\mathrm{L}(x;|F|) = \frac{P(|F|;x)}{P(|F|)} \tag{2.69}$$

is the likelihood function, where $P(|F|;x)$ is the conditional distribution of experi-

mental data when the parameters are known. For x $\mathbf{F}_{calc}$ may be used, since $\mathbf{F}_{calc}$ can be calculated directly from x. Thus

$$P(|F|;x) \equiv P(|F_{obs}|^{all\ reflections};\mathbf{F}_{calc}^{all\ reflections}) \tag{2.70}$$

This joint conditional probability function may be built up by generating the conditional probability distribution of each reflection and multiplying these together:

$$P(|F_{obs}|^{all\ reflections};\mathbf{F}_{calc}^{all\ reflections}) = \prod_{all\ reflections} P(|F_{obs}|;\mathbf{F}_{calc}) \tag{2.71}$$

The refinement consists of maximising the likelihood function, which is proportional to the product function of the conditional probability distributions for each reflection (see equation 2.71). Maximising this function is equivalent to the minimisation of its negative logarithm, giving rise to what is called the '−log-likelihood':

$$-\ln P(x;|F_{obs}|^{all\ reflections}) = -\ln q(x) - \sum_{all\ reflections} \ln \mathrm{L}(F_{calc};|F_{obs}|) \tag{2.72}$$

Because $P(|F|)$ (see equation 2.69) will not change, effectively the function to be maximised is $P(|F_{obs}|;\mathbf{F}_{calc})$.

The principle of maximum likelihood may be formulated as follows: "an organised search for those combinations of phases associated with a 'basis set' of reflections which have maximum likelihood, *i.e.* which lead to the assignment of the highest conditional probability to the observed moduli belonging to reflections outside the basis set" (Bricogne and Gilmore, 1990), as described in the previous section. This is only one of many ways of formulating the use of Bayesian statistics and the principle of maximum likelihood in the field of protein crystallography.

One of the parameters of the likelihood function used in REFMAC is σ_A, which was first described by Srinivasan and Ramachandran (1965). σ_A is a combined measure of the completeness and the accuracy of a partial structure (Read, 1986).

In REFMAC σ_A is estimated from the 'free' reflections (Brünger, 1992) rather than all reflections as by Read, which can be done satisfactorily using only 200 reflections (Murshudov *et al.*, 1996b). From the theories by Luzzati (1952) and Read(1986,1990) the parameters m and D_j are derived:

$$D_j = \langle cos(2\pi \vec{h} \cdot \Delta \vec{r}_j) \rangle \tag{2.73}$$

where $\langle\rangle$ denotes an expected value or probability weighted average. $\vec{h}$ is the reciprocal lattice vector, and $\vec{r}_j$ are the atomic coordinates (in Å).

$$m = \langle cos(\alpha_N - \alpha_P^{calc}) \rangle \tag{2.74}$$

m has different values for acentric and centric reflections, and also depends on whether the partial structure P is known to contain errors or not. $N = P + Q$ is the total number of atoms when Q is the missing structure. Therefore

$$\mathbf{F}_N = |F_N| exp(i\alpha_N) = \mathbf{F}_P^{calc} + \mathbf{F}_Q \tag{2.75}$$

σ_A is then

$$\sigma_A = D_j \sqrt{\frac{\sum_P}{\sum_N}} \tag{2.76}$$

where $\sum_N = \sum_{j=1}^{N} f_j^2 = \langle |F_N|^2/\epsilon \rangle$, in which ϵ is a correction factor for the expected intensity in the appropriate reciprocal-lattice zone, denoting the number of identical contributions arising from symmetry ($\sum_P$ analogous to $\sum_N$). f_j is the atomic scattering factor. σ_A has a value between 0 and 1, where 0 indicates that the partial structure provides no phase information (because it contains no atoms or the partial structure bears no relation to the total structure), and 1 indicates a perfect and complete partial structure. D_j varies with resolution, so σ_A needs to be estimated for several resolution ranges.

The likelihood function in REFMAC has the following form:

$$LLK = \sum_h LLK_h \tag{2.77}$$

where

$$-LLK_h = \begin{cases} c - \ln(|F_{obs}|) + \ln(2\sigma^2_{F_{obs}} + \Sigma_{wc}) + \frac{|F_{obs}|^2+|F_{wc}|^2}{2\sigma^2_{F_{obs}}+\Sigma_{wc}} - \ln I_0(\frac{2|F_{obs}||F_{wc}|}{2\sigma^2_{F_{obs}}+\Sigma_{wc}}) & \text{(acentric)} \\ \frac{1}{2}\ln(\sigma^2_{F_{obs}} + \Sigma_{wc}) + \frac{|F_{obs}|^2+|F_{wc}|^2}{2(\sigma^2_{F_{obs}}+\Sigma_{wc})} - \ln\cosh(\frac{|F_{obs}||F_{wc}|}{(\sigma^2_{F_{obs}}+\Sigma_{wc})}) & \text{(centric)} \end{cases} \tag{2.78}$$

$\mathbf{F}_{wc} = \sum_{j=1}^{N_{part}} D_j \mathbf{F}_j^{calc} = |F_{wc}|exp(i\phi_{wc})$ is the weighted sum of partial calculated structure factors. $\sum_{wc} = \epsilon \sum_{j=1}^{N_{part}} \sum_j (1 - D_j^2)$ where $\sum_j = \sum_{k=1}^{N_{atom}} f_{kj}^2$ for j-th partial structure. N_{atom} is the number of atoms, and N_{part} is the number of partial structures. $\sigma_{F_{obs}}$ is the experimental uncertainty in the amplitude $|F_{obs}|$. I_0 is a zero order modified Bessel function.

Murshudov *et al.* (1996b) describe a number of special cases, where the likelihood function is reduced to simpler functions because of various assumptions and approximations:

1. if some atomic parameters are approximately known and it is assumed that the missing atoms are distributed uniformly over the asymmetric unit the total distribution reduces to a distribution which neglects experimental errors for the missing structure (*i.e.* D_j for the missing structure is 0, and only the partial structure affects D_j) (Srinivasan and Ramachandran, 1965);

2. if the partial structure may be assumed correct and the only error is due to missing atoms, in which case $D_j = 1$ and $\sum_{wc} = \sum_Q$ (Sim, 1959);

3. in the case of a Wilson distribution (*vide supra*) no atomic parameters are known and the contents of the unit cell is distributed uniformly;

4. when $\frac{2|F_{obs}||F_{wc}|}{\sum_{wc}}$ is small the $-$log-likelihood residual is similar to the normalised

intensity based Patterson correlation function (which is related to normalised intensity based least squares) (Bricogne, 1992);

5. when $|F_{obs}| \approx |F_{wc}|$ so that $\ln \frac{|F_{obs}|}{|F_{wc}|}$ is small, the $-$LLK residual is similar to an amplitude based least squares residual, the Gaussian based likelihood function. This assumption is reasonable near the end of refinement;

6. if all reflections are measured with equal $\sigma_{F_{obs}}$ (which is very unrealistic) and all $D_j \approx 1$, $-$LLK will become a unit weighted least squares residual.

From parameters m and D_j Fourier coefficients are generated which are used in map calculations. Because absent reflections cause unpredictable noise (see for example Cowtan (1996)) REFMAC approximates them by setting them equal to their expected value (this is done for all reflections, including the 'free' reflections). Therefore the map coefficients are:

$$2FWT = \begin{cases} (2m|F_{obs}| - D_j|F_{calc}|)exp(i\alpha^{calc}) & \text{for reflections present in refinement} \\ D|F_{calc}|exp(i\alpha^{calc}) & \text{for missing data} \end{cases} \tag{2.79}$$

$$\Delta FWT = \begin{cases} (m|F_{obs}| - D_j|F_{calc}|)exp(i\alpha^{calc}) & \text{for reflections present in refinement} \\ 0 & \text{for missing data} \end{cases}$$

The maps produced are more informative and contain less bias than conventional $2|F_{obs}| - |F_{calc}|$ and $|F_{obs}| - |F_{calc}|$ maps. Luzzati (1953) showed that a map with coefficients $|F_{obs}|exp(i\alpha^{calc})$ shows missing atoms at half weight if the structure is almost complete, and slightly lower when more of the structure is missing. Therefore the conventional map with coefficients $(2|F_{obs}| - |F_{calc}|)exp(i\alpha^{calc})$ will bring up missing atoms with almost full weight. It has been determined (Vijayan, 1980) that for different amounts of missing structure a specific value of n in map coefficients $[n|F_{obs}| - (n-1)|F_{calc}|]exp(i\alpha^{calc})$ is most appropriate. Also, Main (1979) showed that $m|F_{obs}|exp(i\alpha^{calc}) \simeq \frac{1}{2}\mathbf{F}_{obs} + \frac{1}{2}\mathbf{F}_{calc}$ so that coefficients which reduce model bias are given by $(2m|F_{obs}| - |F_{calc}|)exp(i\alpha^{calc})$ in the case of a perfect partial structure.

Read (1986) extended this to the case of a partial structure with errors, resulting in the map coefficients $(2m|F_{obs}| - D_j|F_{calc}|)exp(i\alpha^{calc})$ (see above).

2.7 Protein structure validation

The best statistical validation of a protein structure would be if it were to be determined and refined independently by a statistically significant number of experimenters. Not surprisingly, this never happens in protein crystallography because of the time needed to obtain a fully refined protein structure. The correctness and precision of the atomic parameters in the structure will therefore need to be assessed thoroughly, both during and after refinement. It has to be kept in mind that validation through computer programs is only as good as the parameters prespecified as ideal values inside these programs.

Protein refinement is inherently difficult because of the data being weak, not highly overdetermined, not to atomic resolution and prone to data error. Data quality has already significantly improved during the past five years for various reasons (Dodson, Kleywegt and Wilson, 1996):

- two-dimensional detectors are now routinely used, reducing the problem of systematic errors;
- cryocrystallographic techniques have advanced to a point where crystal quality stays constant during a complete experiment, often even extending the resolution that may be obtained;
- synchrotron radiation allows the use of smaller crystals, while higher resolution data may be obtained and absorption effects are reduced because of shorter wavelengths;
- data processing techniques are advancing, incorporating information on experimental standard deviations.

A first indication of the reliability of a protein crystal structure is the resolution to which the data has been collected and the structure refined. This can only be a rough measure, but it has been shown that at higher resolution the structure is closer to the correct structure (Hubbard and Blundell, 1987). A measure of the accuracy and usefulness of intensity data, by measuring the agreements of equivalent reflections, is:

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^{N} |\langle I_{hkl} \rangle - I_{hkl}(i)|}{\sum_{hkl} \sum_{i=1}^{N} I_{hkl}(i)} \tag{2.80}$$

where $I_{hkl}(i)$ is the i-th measurement of the reflection with indices hkl and $\langle I_{hkl} \rangle$ is the mean value of the N equivalent reflections. In general, R_{merge} values less than 5% indicate excellent data quality, values between 5 and 10% indicate average data quality, values in the 10-15% range are acceptable but may indicate problems with crystals or experiment, while values greater than 15% indicate poor quality data that may not be useful for crystal structure analysis (Ealick, 1995). The value of using as much data as possible has been demonstrated by various people. As mentioned before, Cowtan (1996) has shown how missing reflections affect the reproduction of an image from diffraction data. The completeness of low resolution data is important in the placement of missing parts of the structure, while refinement may benefit from the inclusion of high resolution data even if the merging R is up to 40% (Dodson, Kleywegt and Wilson, 1996). Accurate bond and angle parameters for X-ray protein structure refinement have been extracted from the Cambridge Structural Database and compiled by Engh and Huber (1991), providing a reliable measure for ideal values for bond lengths and angles. Ever increasing computing power allows for more advanced software, both for refinement calculations and graphics facilities. The development of program packages like REFMAC and Quanta's X-AUTOFIT, with extensive maximum likelihood calculations and model building applications respectively, has only been possible because of this.

2.7.1 R-factors

A few specific validation techniques have already been described. The conventional R-factor is the most widely accepted indicator of the general quality of a crystal structure. It is not a good independent validator, since it can be manipulated by excluding data or adding parameters inappropriately. A better indication of the fit between observed and calculated structure factors is the 'free R-factor' (Brünger, 1992), which aids bias-free refinement by indicating overfitting. If the free R-factor does not decrease with a decreasing conventional R-factor in a particular stage of refinement, the new parameters introduced in that stage do not give a significantly improved model. Commonly between 5 and 10 % of diffraction data is excluded from refinement, with a set of between 500 and 1000 reflections considered to produce reliable statistics. Non-crystallographic symmetry will affect the discrepancy between conventional and free R-factors because of phase relationships between reflections in the working and free sets. The expected value of the free R-factor may be estimated as (Cruickshank, 1996):

$$EFRF = R\sqrt{\frac{N_{obs}}{N_{obs} - N_{par}}} \tag{2.81}$$

where N_{obs} is the number of observations, N_{par} is the number of parameters and R is the conventional R-factor.

R-factors as described above are global indicators, which can not point out local errors. Therefore the real space R-factor has been devised (Jones *et al.*, 1991), which can be plotted as a function of the residues along the polypeptide chain:

$$R_{real\ space} = \frac{\sum |\rho_{obs} - \rho_{calc}|}{\sum |\rho_{obs} + \rho_{calc}|} \tag{2.82}$$

ρ_{obs} is taken from a real electron density map calculated from the model on a certain grid G. ρ_{calc} is taken from an 'artificial' electron density map calculated on the same

grid G from a set of idealised Gaussian distribution atoms at all atomic sites in the model, all with average temperature factors. The maps are scaled together with one overall scale factor. A mask is determined for the desired subset of atoms (*i.e.* a residue, or a side chain) on the same grid G, after which the values for ρ_{obs} and ρ_{calc} for this subset may be read from the two map grids, and the R-factor calculated.

2.7.2 Estimates of coordinate precision

Luzzati plot and Cruickshank's σ_d

Luzzati (1952) developed the idea of plots of R'= $\frac{\langle||F|-|F+\Delta F||\rangle}{\langle|F|\rangle}$ (for centrosymmetric structures) and R"= $\frac{\langle||\mathbf{F}|-|\mathbf{F}+\mathbf{\Delta F}||\rangle}{\langle|\mathbf{F}|\rangle}$ (for non-centrosymmetric structures) *vs.* $\sin\theta$, where $\mathbf{\Delta F}$ is the error in structure factor $\mathbf{F}$ due to coordinate errors $\Delta\vec{r}_j$. He argued that the crystallographic R-factor is dependent on resolution (and even on the way the structure factors are calculated), while curves of R' and R" are independent of those factors. They do, however, depend on the assumption that coordinate errors $\Delta\vec{r}_j$ are the sole cause of the difference between $\mathbf{F}_{obs}$ and $\mathbf{F}_{calc}$. R' and R" are functions of $D_j = \langle cos(2\pi\vec{h}\cdot\Delta\vec{r}_j\rangle$ (see equation 2.73) only. If $\Delta\vec{r}_j = 0$ then $D_j = 1$ and R'=R"=0, *i.e.* a completely perfect structure, which would also have the crystallographic R=0. For a completely wrong structure, where the distribution of $\Delta\vec{r}_j$ is constant, D_j tends to 0 and R" (for acentric structures like all natural proteins) approaches $2-\sqrt{2}$. The theoretical curves for R' and R" may indicate an upper limit to the value for $\Delta\vec{r}_j$.

At the Validation Workshop (Dodson, Kleywegt and Wilson, 1996) Cruickshank showed that the Luzzati (1952) plot per se is not an appropriate indication of coordinate error. It is a means for estimating how far the refinement still has to go to reach $R = 0$, assuming that $|F_{obs}|$ has no errors, $\mathbf{F}_{calc}$ is perfect apart from coordinate errors and the Gaussian probability distribution for the coordinate errors is the same for all atoms. The atoms are not required to be identical, and the errors

are not required to be small. However, the assumption of perfection apart from coordinate errors is not applicable. The Luzzati plot may only indicate an upper limit for the coordinate error at the end of refinement. Cruickshank has therefore devised a formula for estimating the expected positional error, caused by diffraction error only, for an atom x with average temperature factor (Cruickshank, 1996):

$$\sigma_d(x) = 0.7[N/p]^{1/2} C^{-1/3} d_{min} R \tag{2.83}$$

where R is the conventional R-factor, N is the number of non-hydrogen atoms in the asymmetric unit, p is the difference between the number of observations and the number of parameters, C is the fractional completeness of the data, and d_{min} is the resolution. The smaller d_{min} and R, the better the precision of the structure. There is a definite direct link between coordinate precision and temperature factor (Dodson, Kleywegt and Wilson, 1996). With ever higher resolution data it is already becoming possible to use least squares matrix inversion at the end of refinement to obtain estimated standard deviations for the parameters describing the well ordered parts of the cell. This will allow testing of Cruickshank's σ_d.

Luzzati, Read and REFMAC: σ_A

Another estimate of coordinate error is σ_A, which has been discussed in section 2.6. Luzzati showed that

$$\begin{aligned} D_j &= \langle cos(2\pi \vec{h} \cdot \Delta\vec{r}_j) \rangle \\ &= exp[-\pi^3(\langle|\Delta\vec{r}_j|\rangle)^2 (sin\theta/\lambda)^2] \end{aligned} \tag{2.84}$$

where $\langle\rangle$ denotes average value. Thus, for $\sigma_A = D_j(\sum_P / \sum_N)^{1/2}$ as discussed before, $\ln \sigma_A = \frac{1}{2} \ln(\sum_P / \sum_N) - \pi^3(\langle|\Delta\vec{r}_j|\rangle)^2 (\frac{sin\theta}{\lambda})^2$. If σ_A can be estimated, a plot of $\ln \sigma_A$ *vs.* $(sin\theta/\lambda)^2$ will yield $|\Delta\vec{r}_j|^2$ from the slope and $(\sum_P / \sum_N)$ from the intercept. However, like for the Luzzati plot, the σ_A theory assumes no errors ex-

cept coordinate errors, so the σ_A-plot needs to be interpreted with care. Brünger has shown that the cross-validated, *i.e.* free R-factor based, Luzzati and σ_A plots correlate better with the actual coordinate errors than those based on conventional methods (Brünger, 1996). The σ_A indicator is used in the calculation of several parameters in REFMAC and thus used explicitly in the improvement of refinement. Maximum likelihood itself is a validation technique in that it uses proper weighting of prior and experimental information by separating errors due to a poor model from experimental errors.

2.7.3 Stereochemistry and residue environment checks

Ramachandran plot

After experiment duplication, the best validation is by criteria that can not and/or have not been used during the course of refinement. One of the validation tools that is very hard to use during automatic refinement, is the Ramachandran plot (Ramachandran *et al.*, 1963). It is therefore a good check of the structure, and residues in odd positions in the plot need further investigation. This does mean that most experimenters now use the Ramachandran plot in the model building stages of structure refinement, so that it is not a completely independent check of the stereochemistry of the final structure. In X-AUTOFIT (Oldfield, 1996) it is now possible to apply weak restraints so that residues will assume the Ramachandran conformation of either a helix, a β-strand, or the nearest allowed region. The default, however, is not to apply any restraints automatically.

SQUID

SQUID displays and analyses molecular coordinates from crystallography, NMR and molecular dynamics (Oldfield, 1992). The structure analysis consists of various atomic checks (*e.g.* distance, chirality, neighbours, non-bond clashes etc.), analysis

of local structure by packing angles, Ramachandran plot and other properties, an overall check of weight, volume, secondary structure, hydrogen bonds and temperature factors, and the possibility of comparison of all those properties with those in other structures. The program outputs detailed warnings about errors and anomalies in the structure under investigation. It provides better independent checks of the structure through root mean square deviations of χ and ω angles, which are not easily tampered with by the experimenter. The mean and spread of the distributions of these angles have been determined from a set of reliable structures.

PROCHECK

PROCHECK is a program to check the stereochemical quality of protein structures (Laskowski *et al.*, 1993). It outputs various plots and a comprehensive residue-by-residue listing, providing an assessment of the overall quality of the structure as compared with well refined structures of the same resolution and highlighting regions which may need further investigation. It may therefore be used to follow the process of refinement and assess the final structure, as well as in the assessment of existing structures to be used for modeling or molecular replacement study purposes. Multiple X-ray crystal structures as well as ensembles derived by NMR spectroscopy can be checked simultaneously.

WHAT-IF and other geometry checking programs

WHAT-IF is a computer program originally written to aid macromolecular modeling and drug design (Vriend, 1990). It provides for displaying, manipulating and analysing small molecules, proteins, nucleic acids, and their interactions, and incorporates access to protein structure databases for reference. WHAT-IF's protein validation portion, called WHAT-CHECK, checks for clashes between symmetry-related molecules, and includes the determination of a 'quality factor' assessing the distribution of different atom types in the environment around side chain fragments.

The expected distributions are compiled from a data set of high resolutions structures (Vriend and Sander, 1993).

Several smaller program packages have been specifically written for protein structure validation. The program PAP (Callahan *et al.*, 1990) provides several plots of main chain geometry, and information on temperature factors. GEOM (Cohen, 1993) provides an analysis of main chain distances, angles and torsion angles, and of side chain torsion angles. ERRAT (Colovos and Yeates, 1993) analyses the relative frequencies of non-bonded interactions between any C, N and O atoms. Because this results in the consideration of only six different pairs, the statistics are better but the information is less specific than in the case of analysing all different atoms specifically, as in *e.g.* WHAT-IF. The program SurVol (Wodak *et al.*, 1995) calculates for each atom in the structure the atomic volume and compares it with a pre-computed average for each atom type. A scoring method was devised in which a high score indicates uncertainty in the structure.

Folding profile methods

Determining the potential fold of a protein sequence is most often approached by searching databases for sequences of known structure that are similar to the sequence under study. This only works if the sequence identity is reasonably high. The advances of the 'tertiary template method' (Ponder and Richards, 1987) were limited because of the rigidity of the methods employed. Eisenberg and co-workers have devised a method which allows more flexibility, both of the backbone of the protein and of the spacing between particular residues (Bowie *et al.*, 1991). Three features determine a residue's environment:

1. the total area of the side chain buried by other protein atoms;
2. the fraction of the side chain area covered by polar atoms or water;
3. the local secondary structure, *i.e.* α-helix, β-sheet and 'other'.

Each residue is then assigned one of 18 environment classes, converting the three-dimensional structure to a one-dimensional string.

It is well established that each of the 20 natural amino acid types has a clear preference for a certain environment. On the basis of well-refined three-dimensional structures a score is assigned to the occurrence of every amino acid in each of the 18 environment classes. These scores can then be used to find the best alignment of amino acid sequences to the one-dimensional string of the protein under study, producing a compatibility score of the sequence with its three-dimensional structure. Local scores may be calculated as an average over 21 residues with the residue-to-be-scored in the centre of the 21 residue window.

Instead of stereochemical parameters, Sippl (1993) uses knowledge of the forces which stabilise proteins in solution in the analysis of the energy distribution of a protein profile. These energy forces are obtained from well-refined structures in the form of potentials as a function of the spatial separation of two atoms. Cα-Cα pairs and Cβ-Cβ pairs seem to work equally well, indicating that the method can be applied if only a Cα-trace is available. The atom pair interaction energy is a function of the sequence, and the structure with the correct conformation will have a lower energy than any alternative conformation. Since no stereochemical parameters are used in the method, the occurrence of high energies in a particular structure is not a consequence of violations of basic steric requirements.

2.7.4 Brookhaven Protein Data Bank

Over the last twenty years, the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) has grown to be a collection of protein and nucleic acid structural data from various sources. Most scientific journals require deposition of atomic coordinates (and often also structure factor data) at the Data Bank upon publication of articles about the structure of proteins, nucleic acids and complexes. Although coordinate checking software is available at the Data Bank and rigorous checks are performed, it is the

ultimate responsibility of the experimenters to provide a reliable set of data pertaining to the structure described in an article. This data may be used for molecular replacement studies, molecular modelling studies etc., and it should be clear from the combination of the article and the Data Bank information, what the limitations are of the information provided.

Chapter 3

Crystallographic studies of seven modified insulins

3.1 Introduction to modified insulins

As described in chapter 1, several features are found in every insulin structure known so far: the three disulfide bridges, the three turn α-helix of residues B9 to B19 and the two small α-helices in the A chain spanning residues A1 to A8 and A13 to A19, together with the short β-sheet structure formed in the dimer, are basic building blocks of insulin (see figure 3.1).

The largest structural differences between insulin monomers in various crystal forms are found in residues B1 to B8. In 2Zn insulin all six monomers in the hexamer have B1 to B8 extended, thus exhibiting the so-called T6-state (for an explanation of the nomenclature of insulin, see chapter 4). In the 4Zn insulin hexamer, B1 to B8 in one of the two molecules becomes helical, giving rise to the T3R3-state. By analogy, it is thought possible to produce a 6Zn insulin in the R6-state. Phenol monoclinic insulin crystals can be obtained in this R6-state, but do not have the expected six off-axial zinc sites, since these are occupied by the phenol molecules.

Crystallographic studies have been carried out on several modified insulins

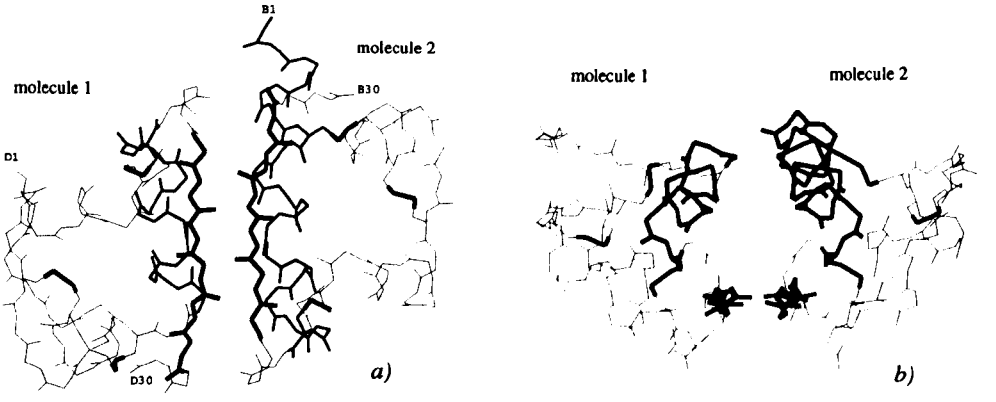

Figure 3.1: The insulin dimer, 4Zn native rhombohedral structure.
a) View along the local twofold axis. Disulfide bridges and dimer interface residues B23-B28 from both molecules in thick lines;
b) View perpendicular to the local twofold axis. Disulfide bridges and residues B23-B28 as before. B chain helices also highlighted in thick lines: D9-D19 (molecule 1) and B1-B19 (molecule 2).

as part of an on-going research programme into the structure of insulin in all its aspects, partly in collaboration with Novo Nordisk A/S who supplied the protein material. Modified insulins are studied structurally to understand for example the conformation of specific parts of the protein, the aggregation of insulin to dimers and hexamers, and metal binding. In the tables and text of this and the following chapters, abbreviated names will be used for the modified insulins, according to table 3.1.

3.1.1 B13 Glu→Gln human insulin

In 2Zn insulin, six glutamic acid residues are brought into close contact in the centre of the hexamer. Electrostatic interactions between those negatively charged residues are unfavourable and are overcome by the zinc binding of the B10 histidine residues, by hydrogen bonds and probably by other cations (Emdin *et al.*, 1980,

name	full description
B13Q	B13 Glu→Gln HI
B9D/B27E	B9 Ser→Asp, B27 Thr→Glu HI
A21G/B9E/B10E	A21 Asn→Gly, B9 Ser→Glu, B10 His→Glu HI
CoI	Zn^{2+}→Co^{2+} HI with iodide
B9H	B9 Ser→His HI
B8S/B13Q/B30amide	B8 Gly→Ser, B13 Glu→Gln, B30 Thr-NH_2 HI
B25Y(B29-A1)A4Q	A4 Glu→Gln, B25 Phe→Tyr, des† B30 single chain HI

Table 3.1: Abbreviations for names of the modified insulins.
HI = human insulin; †des = deleted

Hill *et al.*, 1991). The B13Q mutant is expected to associate specifically into hexamers, stabilised by hydrogen bonds between the glutamine residues in a twelve-membered 'ring' structure in the core as proposed by Markussen *et al.* (1987), shown in figure 3.2.

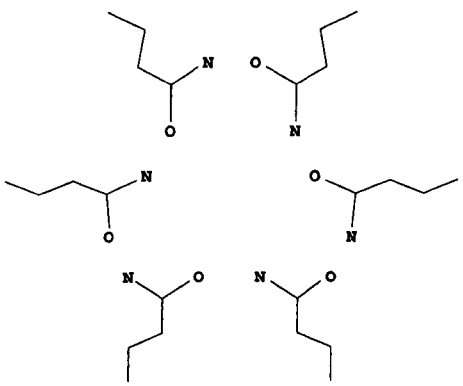

Figure 3.2: The proposed 'ring' structure of B13 glutamine residues in B13Q insulin.

Because Gln at position B13 favours the formation of helix at B1 to B8 (Bentley *et al.*, 1992), this more stable hexamer is also expected to have undergone the T → R structural transition, especially upon binding of zinc: the zinc free, 2Zn (both studied by Xiao and co-workers (Bentley *et al.*, 1992)) and 4Zn forms

of the B13Q mutant should progressively prefer the R-conformation. According to Markussen *et al.* (1987) B13Q has a potency relative to insulin of 0.18 in the free fat cell assay.

3.1.2 B9 Ser→Asp/B27 Thr→Glu human insulin

Introduction of negative charges at B9 in what is the core of the 2Zn native insulin hexamer will prevent hexamer and dimer formation. The B9 Ser→Asp mutation is one of the most effective in reducing dimer formation at neutral pH (Brange *et al.*, 1988). In a dimer, two B9 residues and two B13 residues (all carboxylate groups in the mutant) would come very close together, causing strong electrostatic repulsion (Dodson *et al.*, 1993), which can be alleviated by protonation and hydrogen bonding. In that case the B13 Glu has an elevated pK, as determined by Kaarsholm *et al.* (1990). At neutral pH the B13 Glu would be deprotonated, in which case hydrogen bonding can not take place and the monomers can not approach each other close enough to form dimers. Dimerisation will also be impaired by changing the interactions of B27 and B28 with the termini of the A chain. The mutant is indeed found to be monomeric at 10^{-3}M (Brange *et al.*, 1988). The structure of the monoclinic form of the mutant has been determined by J. Turkenburg (1992). In that case the mutant forms hexamers with calcium ions bound in the core which neutralise the B9 mutation.

3.1.3 A21 Asn→Gly/B9 Ser→Glu/B10 His→Glu human insulin

Addition of a negative charge at B9 introduces electrostatic repulsion and will prevent hexamer and dimer formation as discussed in section 3.1.2. In addition, abolition of the metal binding potential at B10 is expected to obliterate hexamerisation, promoting rapid absorption after subcutaneous injection. Substitution of A21 Asn

by Gly is expected to decrease susceptibility to deamidation and covalent dimer formation, thus stabilising the mutant in the, often acidic, therapeutic solutions. The rate of absorption of some A21 substituted insulin mutants from the injection site appears to be virtually independent of zinc concentrations. The A21G/B9E/B10E mutant is found to be monomeric under physiological conditions (Xiao, 1990). The relative biological activity (mouse adipocyte assay) is 58% (N. Kaarsholm, pers. comm.). This value is composed of a significantly enhanced activity afforded by the B10 Glu substitution, a decreased activity provided by the B9 Glu mutation, and a slightly decreased effect on activity from the A21 Gly substitution. In acid-base titrations to characterise ionisation patterns of mutant insulins, it is found that carboxylate groups residing close to the monomer-monomer interface have their pK values perturbed upwards (Kaarsholm *et al.*, 1990). Unfortunately, no data are available for an acidic side chain at B10. However, the situation will be further complicated in that case, because of the potential for this side chain to interact with the B9-B19 helix dipole (which would tend to lower pK) adding to the close proximity of the monomer-monomer interface (which would effect an increased pK as observed with carboxylates at B13 and B9) (N. Kaarsholm, pers. comm.).

3.1.4 Co^{2+} human insulin

Cobalt is very similar to zinc in size (ionic radii 0.72Å and 0.88Å respectively (Weast, 1983-1984)) and chemical behaviour, and is therefore probably able to play the same role in insulin hexamer formation. In native insulin hexamers, zinc is present either in octahedral conformation, when the insulin molecule is in the T-state, or in tetrahedral conformation, when the insulin molecule is in the R-state. According to ligand field theory, however, tetrahedral coordination is less favourable than octahedral coordination for d^7 ions like Co^{2+}. In the presence of iodide (and other halide ions), tetrahedral coordination is a possibility for Co^{2+} (Mackay and Mackay, 1981). The addition of phenol in the crystallisation is ex-

pected to stabilise α-helix at B1-B8, favouring R- over T-conformation (Derewenda *et al.*, 1989), which probably forces Co^{2+} into a tetrahedral environment. Whether a Co^{2+} structure has T3R3- or R6-conformation depends on the phenol concentration (Whittingham *et al.*, 1995) and the temperature of the crystallisation (de Graaff *et al.*, 1981).

3.1.5 B9 Ser→His human insulin

Introduction of another histidine residue into the core of the insulin hexamer is expected to add to the stability by increased zinc binding. In fact, solution studies at pH 8 have shown the B9H mutant is capable of binding around 6 zinc ions per insulin hexamer. As it turns out, this additional stability is not sufficient to provide a clinically relevant protracted profile following subcutaneous injection (N. Kaarsholm, pers. comm.). This mutant could, in rhombohedral crystal form, be an example of a 6Zn insulin in the R6-state, with six off-axial zinc binding sites. In order to promote T → R structural transition to facilitate the formation of the R6-state, phenol was added in an attempt to obtain a different crystal form.

3.1.6 B8 Gly→Ser/B13 Glu→Gln/B30 Thr-NH_2 human insulin

Introduction of positive charges at B13 and B30, which raises the iso-electric point, is expected to create insulins more soluble in acidic therapeutic solutions. At the higher pH in the body, they would crystallise readily after injection, resulting in prolonged action. A charge-indifferent mutation at B8 is introduced to study the background of the sequence invariance at that point, which might be related to the conformation of the B8 residue in the native molecule. The hydrophobicity of the mutant is increased in comparison with native pig insulin as determined by HPLC chromatography (Markussen *et al.*, 1987). This is unexpected because of

the hydrophilic nature of the mutation at B8 as such. For the T-conformation of the mutant Markussen proposes a distortion of the α-helix commencing at B7 (to accommodate an L-Ser residue rather than a Gly with D-amino acid Ramachandran angles as in the native molecule (Baker *et al.*, 1988)), which might expose hydrophobic side chains. The glycine at B8 in native insulin in the R-conformation, however, has Ramachandran angles as expected for an α-helix. The introduction of serine in that position might in this conformation not have such a significant effect on the helix. Phenol should stabilise α-helix at B1-B8, like in monoclinic native insulin (Derewenda *et al.*, 1989).

3.1.7 A4 Glu→Gln/B25 Phe→Tyr/des B30 single chain human insulin

A4 Glu is involved in protamine binding, which retards the uptake of insulin in the bloodstream and is used in some therapeutic mixtures (J. Brange, pers. comm.). Changing this residue from a negatively charged polar Glu to a neutral Gln is expected to reduce repulsion within the hexamer and generate more effective association. This mutation has a profound effect on the position of B25 (Holden, 1991) and its interaction with A19 Tyr (Xiao, 1990), which for the 2Zn native molecule is shown in figure 3.3. The interaction between B25 and A19 in T-conformation insulin is one of the few hydrogen bond interactions between the two chains, and thus naturally was thought to be important for the stability of the molecule.

Des B30 insulins have biological activities very similar to full length insulins (Baker *et al.*, 1988). However, introduction of a peptide link, causes nearly complete abolition of activity (Derewenda *et al.*, 1991), attributed to the rigidity of the molecule, and therefore the impossibility to achieve the conformational change probably required for binding to the receptor. B25Y(B29-A1)A4Q is a single chain precursor for an insulin analogue designed for the study of the influence of the biological activity of amino acid substitutions at position B25. The B25Y(B29-A1)A4Q

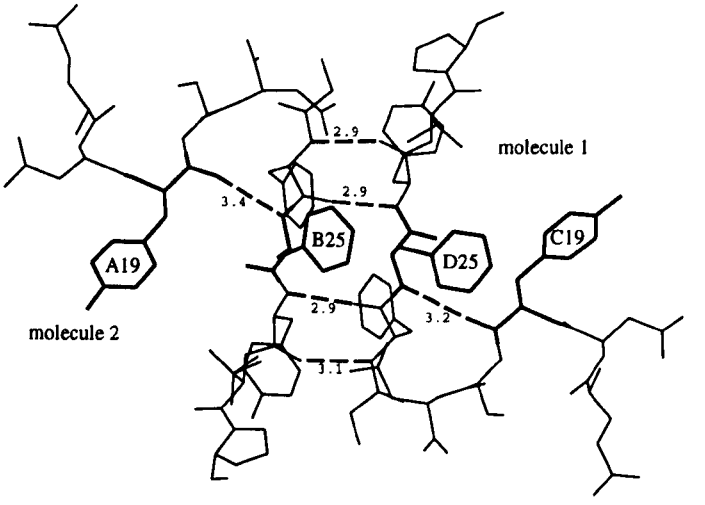

Figure 3.3: The interaction of B25 with A19 in 2Zn native insulin. B25 and A19 residues from both molecules highlighted in thick lines. Distances in Å.

mutant is found to react only very slowly with trypsin at the B29 Lys residue, while this reaction is a key step in the final preparation of the desired insulin analogue (P. Balschmidt, pers. comm.).

3.2 Crystallisation of the modified insulins

The routines for insulin crystallisation in various crystal forms had become well established by 1975 (Cutfield, 1975). These routines were reviewed by Tolley (1987) and improved and adapted by Xiao (1990). The present standard routines for the various crystal forms of native insulin can be found in table 3.2.

Most of the insulins described in this thesis have been crystallised by batch method by various people. The actual conditions employed in the crystallisation of the insulins by batch method can be found in table 3.3. Two mutants have been crystallised by vapour diffusion with the following protocols (both by J. Clarkson):

additive	rhombohedral 2Zn†	rhombohedral 4Zn‡	monoclinic¶	orthorhombic†	cubic‡
insulin	6 mg/ml	4 mg/ml	10 mg	3-6 mg	5 mg/ml
0.02M HCl	2.5 ml	1.0 ml	2.0 ml		
0.05M HCl					0.5 ml
0.005M H_2SO_4				1.0 ml	
ZnAc	10 mg				
0.15M ZnAc		0.15 ml			
0.12M ZnAc			0.1 ml		
0.05M Na_2SO_4				1.0 ml	
acetone	1.0 ml	0.3 ml			
0.2M Na_3citrate	1.25 ml	0.5 ml	0.5 ml		
0.82M NaCl in 0.0136M EDTA					0.88 ml
5%aq phenol			0.4 ml		
phenol, m-cresol or xylene					0.3%
salt§		9%	$\leq$ 120 mg		
H_2O	0.25 ml				
pH	6.5	6.2	6.5 - 7.8	2.4 - 3.7	9.3

Table 3.2: Standard routines for crystallisation of insulin in various crystal forms. Adapted from Tolley (1987) and Xiao (1990).
†from Xiao (1990); ‡from Tolley (1987) with adaptations by Xiao (1990); ¶from Tolley(1987); §NaCl, NaBr, NaI or NaSCN

	B13Q	A21G/B9E/B10E	B8S/B13Q/B30amide	CoI	B9H -1	B9H -2
insulin	4 mg/ml	4mg	3 mg	75 mg zinc free	4.8 mg	3.27 mg
HCl		0.52 ml 0.02M	0.4 ml 0.02M	10 ml 0.02M	560 μl 0.02M	600 μl 0.02M
NaCl	9%		12 mg		61.0 mg	20.9 mg
NaI				450 mg		
Na_3cit	0.04M	0.25 ml 0.2M	100 μl 0.2M	2.5 ml 0.2M	250 μl 0.2M	160 μl 0.2M
acetone	16% v/v	0.2 ml				
ethanol					200 μl	
$ZnAc_2$	7.5 mM	10 μl (100mg/ml)	20 μl 0.2M		1.7 mg	30 μl 0.15M
$CoAc_2$				0.5 ml 0.12M		
phenol			80 μl 5%aq	2.0 ml 5%aq		130 μl 5%aq
pH	6.4	6.57	6.60	6.29	7.43	7.27
temp (°C)	65	20	55	55	55	54
result†	rh	or	mc	rh	rh	mc

Table 3.3: Crystallisation protocols of the batch crystallisation of five modified insulins, one of which in two different crystal forms.
†resulting crystal form: rh=rhombohedral; or=orthorhombic; mc=monoclinic

1. B9D/B27E:

   - Add 10 mg insulin to 200 μl 0.02 M HCl and 8 μl 0.5 M NaOH. The mixture precipitates. Store in refrigerator for 72 hours;
   - Add 100 μl 0.02 M HCl to partly dissolve the protein. Centrifuge and use supernatant for crystallisation by vapour diffusion in hanging drop. Store surplus supernatant in refrigerator for 24 hours;
   - Supernatant precipitates. Warm up (45 °C) to partly dissolve the protein. Centrifuge and use supernatant for crystallisation by vapour diffusion in hanging drop. The reservoir solution contained 1 ml 0.02 M citrate/phosphate buffer pH6. The drop solution contained 4 μl of buffer/protein solution in a 40:60 ratio. This protocol resulted in orthorhombic crystals suitable for diffraction experiments.

2. B25Y(B29-A1)A4Q: The reservoir solution was made up of 995 μl 0.02 M citrate/phosphate buffer pH6 and 5 μl saturated ammonium sulphate. The drops contained 5μl of a solution of 9.9 mg insulin in 200 μl 0.02 M HCl. This protocol resulted in cubic crystals under a brown skin, due to the impurities in the insulin preparation. However, crystallisation attempts after purification failed. Careful analysis of the original insulin preparation by Novo Nordisk A/S revealed the presence of 0.25 zinc ion per monomer insulin (P. Balschmidt, pers. comm.), which would have been removed by the purification steps. Zinc was obviously a necessity in the successful crystallisation of the mutant.

Further crystallisation experiments include:

- The batch crystallisation of B8S/B13Q/B30amide in cubic crystal form in order to study the conformation of B8 without the constraints of metal binding and hexamerisation. The crystals obtained were unsuitable for diffraction experiments;

- The batch crystallisation of B9H in rhombohedral crystal form in the presence of glycerol instead of ethanol. This would facilitate cryocrystallographic experiments. Crystals of rhombohedral B9H grown under the standard conditions (see table 3.3) deteriorated quickly upon addition of the cryoprotectant. Crystallisation experiments under the new conditions were unsuccessful. Brange and others (see Brange, 1987) have observed that the addition of glycerol to a neutral solution of insulin impairs the chemical stability of insulin. Glycerol is, however, also used as an isotonic agent in some pharmaceutical insulin preparations;

- In a separate attempt to obtain B9H crystals suitable for cryocrystallographic experiments, a cryoprotectant solution was made with all the constituents of the standard mother liquor apart from insulin, and 1,2-ethanediol instead of ethanol to function as cryoprotectant. 346 μl aliquots of a standard solution containing 1250 μl 0.2M Na_3Cit, 8.3 mg $ZnAc_2$, 304.5 mg NaCl, 85 μl 1M NaOH, 35μl 1M HCl and 280 μl 0.2M HCl, mimicking the original crystallisation solution, were subsequently diluted with varying amounts of ethanediol and water in order to obtain a suitable cryoprotectant solution. Crystals from the original rhombohedral crystallisation were submerged in a solution with 30% cryoprotectant, to check their viability in solutions with ethanediol. The crystals did not change noticeably over a period of 24 hours. The 30% cryoprotectant solution was tested separately for glassy appearance upon flash-freezing, and was found to be unsuitable. However, a 40% ethanediol solution froze satisfactorily. The crystals in the 30% solution were taken to the synchrotron at Daresbury (UK), where the solution was replaced with the 40% solution, after which the crystals proved suitable for data collection.

J. Clarkson has attempted to obtain fresh crystals of A21G/B9E/B27E and B25Y(B29-A1)A4Q for data collection at cryogenic temperatures, without success.

insulin	source†	detector‡	temp¶	processed with	reso (Å)
B13Q	rot.an.	Xen	RT	Xengen	1.9
B9D/B27E	rot.an.	Xen	RT	XDS	2.3
A21G/B9E/B10E -1	rot.an.	Rigaku	RT	teXsan	2.8
A21G/B9E/B10E -2	rot.an.	Xen	cryo	XDS	2.3
B25Y(B29-A1)A4Q	DL 9.6 (0.89Å)	Raxis	RT	MOSFLM/DENZO	2.8
B8S/B13Q/B30amide	rot.an.	Raxis	RT	RDPS§	2.1
CoI	rot.an.	Xen	RT	XDS	1.9
B9H rhombohedral -1	rot.an.	Xen	RT	XDS	2.0
B9H rhombohedral -2	DL 9.6 (0.89Å)	Mar	cryo	DENZO	1.9
B9H monoclinic	rot.an.	Raxis	RT	DENZO	2.1

Table 3.4: Data collection information.
†rot.an. = rotating anode; DL = Daresbury Laboratory
‡Xen = Xentronics Area Detector; Raxis = R-AXIS IIC Imaging Plate;
Rigaku = Rigaku AFC5 Diffractometer; Mar = MarResearch Imaging Plate
¶RT = room temperature; cryo = cryogenic temperature, 120K
§RDPS = R-AXIS IIC Data Processing Software for VAX workstations.

3.3 Data collection and processing

Data were collected by various people on various sources and detectors, according to availability of apparatus and requirements of individual projects. The information pertaining to data collection protocols is shown in table 3.4. Statistics of the data used for refinement, are given in table 3.5.

3.4 Structure Determination

The first structure of the insulin molecule has been known since 1969 (Adams *et al.*, 1969). Isomorphous replacement studies of modified insulin are, therefore, normally unnecessary. For modified insulins which seem isomorphous with 2Zn insulin or one of the other forms of insulin determined so far, structure determination and refinement starts with the calculation of Fourier maps with the phases of the isomorph.

	I/sigI	R_{merge}
B13Q	$\geq$3 at 2.2Å	15.77% at 2.4Å, 31.92% at 2.18Å
CoI	88% >3 at 1.9Å	21.4% at 1.88Å
B9H-rh†	3.67 in highest bin (1.92Å)	22.6% in highest bin
B9H-mc‡	72.0% >3 in highest bin (2.1Å)	21.1% in highest bin
B8S/B13Q/B30amide	2.93 between 2.25 and 2.06Å	5.59% at 2.0Å
B9D/B27E	-	-
B25Y(B29-A1)A4Q	70.1% >3 at 2.8Å	7.2% overall
A21G/B9E/B10E	-	-
	completeness	comments
B13Q	84% at 2.284Å	
CoI	91.9% at 1.996Å, 65.2% at 1.88Å	
B9H-rh	94.7% in highest bin	
B9H-mc	99.7% in highest bin	
B8S/B13Q/B30amide	92.8% between 2.25 and 2.06Å	
B9D/B27E	-	not retrievable
B25Y(B29-A1)A4Q	81.9% at 2.8Å	
A21G/B9E/B10E	-	complete to 3.6Å, strong terms to 2.8Å

Table 3.5: Data statistics for the data sets used in the refinements.

Regarding structures described in this thesis, this was possible in most cases, namely B13Q, B9D/B27E, B8S/B13Q/B30amide, CoI, and both crystal forms of B9H. Cell constants, space group and isomorphous R are given in table 3.6 for the modified insulins studied, with the appropriate information for native structures for comparison.

The remaining structures, namely A21G/B9E/B10E and B25Y(B29-A1)A4Q, required complete molecular replacement studies to produce a starting model for refinement. In order to study the symmetry of the aggregation of all eight different structures, self-rotation functions were calculated.

3.4.1 Molecular replacement

B9 Ser→Asp/B27 Thr→Glu human insulin

The molecular replacement studies of B9D/B27E only served to establish the isomorphism of the mutant with orthorhombic native pig insulin. At first, the data were indexed in a cell with a=37.99Å, b=51.54Å, c=57.90Å. A rotational search in *AMoRe* with an orthorhombic native dimer (β_{max}=90°) resulted in a set of 15 peaks

structure*	space group	cell constants†	Z'‡	R_{iso}	isomorph
2Zn native (pig)	R3	a=b=82.500,c=34.000 $\alpha=\beta=90,\gamma=120$	2	-	-
4Zn native (pig)	R3	a=b=80.700,c=37.600 $\alpha=\beta=90,\gamma=120$	2	-	-
monoclinic native (human)	$P2_1$	a=61.230,b=61.650,c=48.050 $\alpha=\gamma=90,\beta=110.50$	6	-	-
orthorhombic native (pig)	$P2_12_12_1$	a=57.979,b=51.534,c=38.003 $\alpha=\beta=\gamma=90$	2	-	-
cubic native (pig)	$I2_13$	a=b=c=78.900 $\alpha=\beta=\gamma=90$	1	-	-
B13Q	R3	a=b=80.815,c=37.591 $\alpha=\beta=90,\gamma=120$	2	0.166	4Zn native
B9D/B27E	$P2_12_12_1$	a=57.90,b=51.54,c=37.99 $\alpha=\beta=\gamma=90$	2	0.452	orthorhombic native
A21G/B9E/B10E	$P2_12_12_1$	a=46.29,b=51.69,c=44.38 $\alpha=\beta=\gamma=90$	2	0.581§	not isomorphous
B25Y(B29-A1)A4Q	$P4_232$	a=b=c=66.78 $\alpha=\beta=\gamma=90$	1	N/A¶	not isomorphous¶
B8S/B13Q/B30amide	$P2_1$	a=60.91,b=62.04,c=47.88 $\alpha=\gamma=90,\beta=110.48$	6	0.192	monoclinic native
CoI	R3	a=b=80.48,c=37.92 $\alpha=\beta=90,\gamma=120$	2	0.208	4Zn native
B9H rhRT	R3	a=b=80.79,c=37.47 $\alpha=\beta=90,\gamma=120$	2	0.261	4Zn native
B9H rhF	R3	a=b=79.878,c=37.427 $\alpha=\beta=90,\gamma=120$	2	0.289	4Zn native
B9H mc	$P2_1$	a=61.14,b=62.09,c=48.08 $\alpha=\gamma=90,\beta=110.43$	6	0.160	monoclinic native

Table 3.6: Crystallographic information of native and modified insulins.

* rhRT = rhombohedral form, room temperature data collection; rhF = rhombohedral form, data collection at 120K; mc = monoclinic form;

† a, b and c in Å; α, β and γ in °;

‡ Z' is the number of molecules in the asymmetric unit;

§ isomorphous R-factor based on orthorhombic native pig insulin structure;

¶ B25Y(B29-A1)A4Q is presumably isomorphous with rat insulin II and insulin from the snake *Zaocys dhumnades dhumnades Cantor* which have been reported to crystallise in $P4_232$ with similar cell constants (Wood *et al.*, 1978, and Liang *et al.*, 1984, respectively), but no coordinates are available for insulin in this space group.

with a correlation coefficient larger than half of that of the first peak. The first four peaks are given in table 3.7.

solution	α	β	γ	cc
1	55.81	84.94	295.26	35.5
2	72.58	59.53	242.83	34.1
3	124.35	90.00	115.29	32.0
4	158.16	38.73	285.12	28.3

Table 3.7: Rotation function solutions for B9D/B27E. cc = correlation coefficient; α,β,γ are Eulerian angles within the *AMoRe* conventions.

solution	α	β	γ	t_x	t_y	t_z	cc	R
1-1	55.81	84.94	295.26	0.0395	0.2180	0.3007	56.2	41.4
1-2	55.81	84.94	295.26	0.0387	0.2851	0.2976	36.0	49.0
2-1	72.58	59.53	242.83	0.4607	0.2829	0.2998	59.1	40.5
2-2	72.58	59.53	242.83	0.4614	0.2196	0.3010	39.5	48.4
3-1	124.35	90.00	115.29	0.4569	0.2172	0.2016	57.2	41.4
3-2	124.35	90.00	115.29	0.4562	0.2175	0.2829	36.3	49.0
4-1	158.16	38.73	285.12	0.1567	0.3133	0.3423	18.4	55.2
4-2	158.16	38.73	285.12	0.0927	0.2279	0.3673	16.2	56.2

Table 3.8: Translation function solutions for B9D/B27E. t_x,t_y,t_z are fractional translations; R = R-factor. Note that the translations for peak 1-1 and 3-1 are the same, apart from whole and half unit cell translations.

The translation search revealed the true nature of these four peaks, shown in table 3.8, each along with their second best translations. The rigid-body refinement clarified them even further, after which the three best solutions, transformed so as to act on the input coordinates, are as follows:

solution	α	β	γ	t_x	t_y	t_z	cc	R
1	359.25	91.92	359.26	-9.77	-12.07	15.13	73.6	35.2
2	180.76	88.07	179.26	28.77	-12.07	13.81	73.6	35.2
3	358.43	33.40	0.67	13.51	-8.78	6.53	69.0	39.1

The first solution indicates identity with just a rotation of 90° around the y-axis, which would put the x-axis along $-z$ and the z-axis along $+x$. This effect can also be achieved by reindexing (h,k,l to l,-k,h), which at the same time transforms the unit cell constants in such a way that they resemble closely the orthorhombic native pig insulin cell constants (see table 3.6). *AMoRe* confirms the validity of reindexing, changing solution 1 to

α	β	γ	t_x	t_y	t_z	cc	R
337.41	1.98	21.88	83.63	37.68	21.50	72.2	35.7

When β=0, only the sum of $\alpha + \gamma$ is relevant, giving the total rotational component of the transformation of the model molecule into the unknown cell (this is a well-known property of Euler angles). Hence in this case identity is achieved by rotating 360°(337.41 + 21.88) around the z-axis. Thus after reindexing the B9D/B27E mutant is isomorphous, judged from the cell constants and the molecular replacement, although the R_{iso} is 0.452. Upon careful comparison of the mutant data with the orthorhombic native data, the relatively high value of R_{iso} can be traced back to a small percentage of the data. According to SCALEIT (CCP4 suite, 1994), only 17 out of the 4548 reflections from the mutant data set have a large difference as compared with those in the native data set, which incidentally was collected on the same instrument.

The direction cosines of the non-crystallographic twofold axis of the refined dimer as determined by overlapping the two monomers, correspond to angles of around (30,90,120), *i.e.* this axis lies in the xz-plane, 30° away from the positive x-axis and 120° away from the positive z-axis. The self-rotation function for the B9D/B27E mutant shows peaks at 37.9% of the maximum height with polar angles (ω, ϕ, κ)=(119.5,0,180), which correspond to the same direction cosines as mentioned before: 0.8703, 0.0000, -0.4925.

A21 Asn→Gly/B9 Ser→Glu/B10 His→Glu human insulin

For mutant A21G/B9E/B10E, the cell constants do not resemble those of any known insulin structure, indicating non-isomorphism. In addition, the R_{iso} is 0.581, against the data of the first choice of isomorph, namely native orthorhombic insulin. The cell volume of the mutant is 6.5% smaller than that of orthorhombic native insulin, indicating the same protein content in slightly closer packing. Molecular replacement studies were done with molecule 2 (chains A and B) from the orthorhombic native dimer as a search model. One peak in the rotation function stood out with a correlation coefficient (cc) of 23.2, where subsequent peaks are 20.9, 19.7, 19.4 etc. The highest 50 peaks of the rotation function were then used in a translation search. The highest peak was very clear, with cc=24.4 and R=51.7, with its second best translation also the second highest peak overall (cc=14.8 and R=54.5). The highest peak was fixed in a translation search for a second monomer, using the next 17 highest peaks from the first translation function (with their translations reset to zero). A second peak was found with a high correlation coefficient (47.1, whereas no other peak was higher than 25.1), with a translation of roughly $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ compared to the first monomer:

	α	β	γ	t_x	t_y	t_z	cc	R
fixed monomer	131.03	57.42	21.74	0.07349	0.38596	0.07461	24.4	51.7
output highest peak	131.03	57.42	21.74	0.57351	0.89027	0.57500	47.1	61.9

Rigid body fitting of these two peaks improved the solution considerably: cc=70.4 and R=38.1 for the two monomers together, and rotations and translations were as follows (internal AMoRe format):

α	β	γ	t_x	t_y	t_z	cc	R
132.04	56.25	25.43	0.06372	0.38815	0.05983	70.4	38.1
139.49	53.72	16.82	0.58117	0.91045	0.57870	70.4	38.1

These monomers are clearly related by a translation of $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ and a small rotation, and thus can not form a dimer directly. The possibility of a symmetry equivalent of one of the monomers complementing the other monomer in dimer formation was investigated. This would necessitate the non-crystallographic twofold axis of the dimer to be parallel to one of the 2_1 screw axes of the space group symmetry. In order to double-check this interpretation, the molecular replacement calculations were repeated with a full orthorhombic dimer as a model. The correlation coefficient of the first peak in the rotation function is distinctly higher than any of the others: 30.2 with the subsequent peaks at 20.9, 20.8, 20.8, 19.3 etc. After the translation function the result is unambiguous: cc=46.0, R=52.0 for the first peak, with cc=14.6, R=54.9 for the second best peak. The highest peak refines to (shifted solution)

α	β	γ	t_x	t_y	t_z	cc	R
176.47	29.68	184.16	85.15	107.73	19.95	67.3	41.2

Investigation of the non-crystallographic symmetry of the refined dimer by overlapping one monomer on the other, produces direction cosines (0.99932, -0.00497, 0.03672), corresponding to angles of almost (0,90,90). This means the non-crystallographic twofold axis is indeed parallel to the x-axis and thus to a 2_1 screw axis. The self-rotation confirms the presence of twofold axes (20.7% of the height of the origin peak) almost along the cell axes, with direction cosines (0.9890, 0.1478, 0.0000).

A4 Glu→Gln, B25 Phe→Tyr, des B30 single chain human insulin

Two insulins have been crystallised in space group $P4_232$, like mutant B25Y(B29-A1)A4Q, namely rat insulin II and insulin from an Asian snake. Both have cell constants very similar to those of the mutant. Atomic coordinates, however, are not

available. The relatively high solvent content of the rat insulin crystals rendered them unsuitable for data collection (Wood *et al.*, 1978). The study of the snake insulin in China was stopped because the supply of the protein was insufficient (Liang D.C., pers. comm.). The sequences of the three insulins are compared in table 3.9. The snake insulin has two significant changes: B2 is proline instead of valine, which puts some restrictions on the conformation of the chain N-terminus, and B5 is arginine instead of histidine, which will prevent alternative zinc binding.

Molecular replacement studies were done with a variety of models: monomers from various crystal forms (2Zn rhombohedral, 4Zn rhombohedral, cubic), dimers and hexamers, either complete or trimmed to the most rigid core by removing chain termini. The most successful model is a 2Zn rhombohedral dimer, with B1-7 removed from both B chains. Hexameric models were mainly used to check the solution and the packing. The rotation function does not show many features. The highest peak has a correlation coefficient of 34.3, and 20 other peaks have a correlation coefficient higher than half that. In the translation function the 20th rotation function peak (which had a *cc* of 22.1) is clearly the highest with *cc*=46.2 and *R*=46.1:

α	β	γ	t_x	t_y	t_z	*cc*	*R*
47.90	90.00	278.79	0.1010	0.5998	0.2476	46.2	46.1

The mutant has one molecule in the asymmetric unit and from packing studies of the dimer solution it was clear that molecule 1 was nearest to the correct solution for the mutant structure. Rigid body fitting was performed on both of the monomers separately. The refined solutions are

α	β	γ	t_x	t_y	t_z	*cc*	*R*
47.53	89.79	278.53	0.1045	0.6010	0.2509	52.8	44.3
48.04	91.41	279.46	0.1108	0.5936	0.2529	52.8	44.3

position	mutant[1]	rat II[2]	snake[3]	occurrence in nature[4]
A4	Gln	Asp	Glu	Glu highly conserved, alternative mainly Asp
A8	Thr	Thr	Glu	mainly Thr (mammals) and His (fish), Glu in some reptiles and birds
A9	Ser	Ser	Asn	mainly Ser in mammals, Asn in most reptiles and birds
A10	Ile	Ile	Thr	mainly Ile (mammals) and Phe (fish), but Thr common in reptiles, amphibians and birds
A15	Gln	Gln	Glu	Gln conserved in mammals, mainly Asp and Asn elsewhere, Glu uncommon
B1	Phe	Phe	Ala	mainly Phe in mammals, but Ala very common in reptiles, birds and fish
B2	Val	Val	Pro	mainly Val (mammals) or Ala elsewhere, but Pro not uncommon in reptiles and fish
B3	Asn	Lys	Asn	mainly Asn in mammals, birds, reptiles and amphibians, Lys only in mouse and rat
B5	His	His	Arg	His highly conserved with Arg as only alternative, all over nature
B16	Tyr	Tyr	Phe	Tyr highly conserved; Phe only seen in snakes
B18	Val	Val	Ile	Val highly conserved; Ile only seen in snakes, Ala in hagfish
B27	Thr	Thr	Ser	mainly Thr (mammals) and Asn (fish), but Ser common in reptiles and birds
B29	Lys	Met	Arg	Lys highly conserved, Arg only seen in snakes, Met only in some rodents
B30	-	Ser	Thr	Thr common in mammals, birds and reptiles; Ser and Ala also common

Table 3.9: Sequence differences between three insulins in space group $P4_232$.

[1] insulin mutant B25Y(B29-A1)A4Q;

[2] insulin type II from rat;

[3] insulin from the Asian snake *Zaocys dhumnades dhumnades Cantor*;

[4] from Derewenda (1990)

The second solution was then applied to a complete molecule 1 from the 2Zn native insulin dimer in the correct cell. The space group symmetry (see figure 3.4) causes six monomers to come together in a T6 hexamer of 32 symmetry.

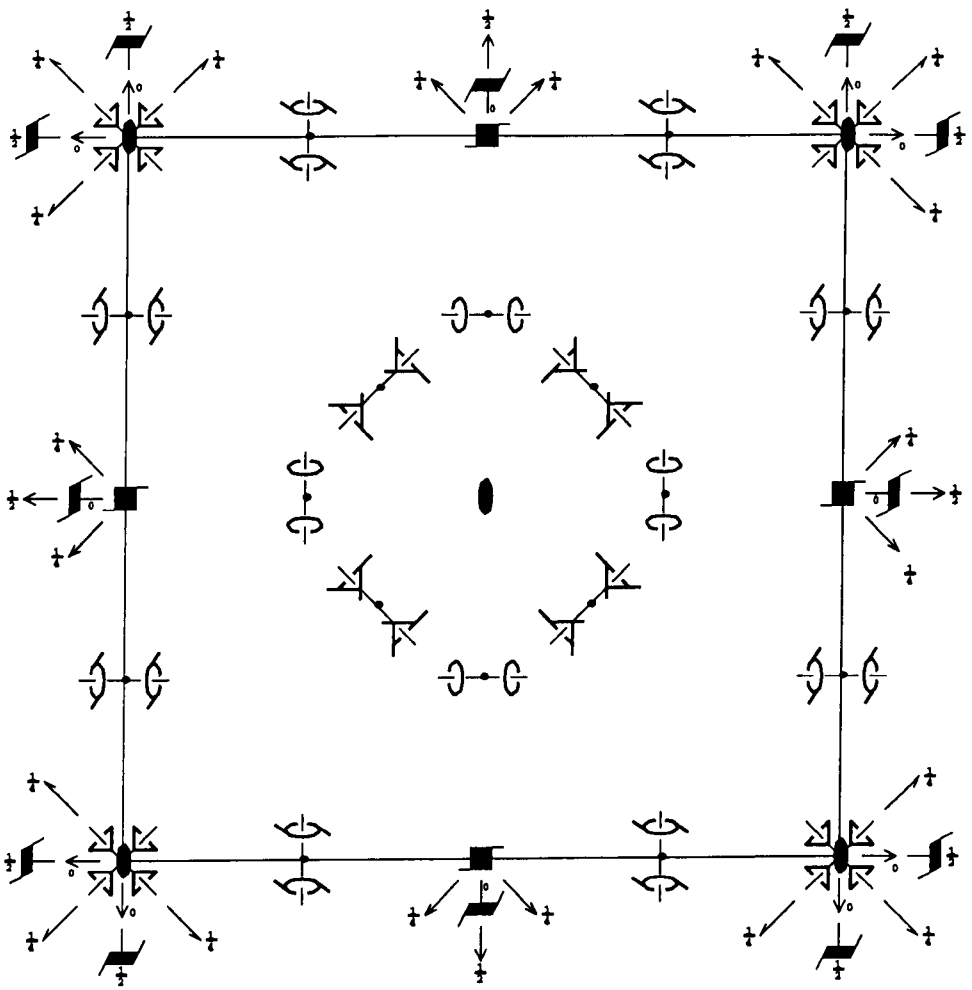

Figure 3.4: Symmetry of space group $P4_232$. Reproduced from the International Tables Volume C (Wilson, 1992).

3.4.2 More self-rotations

B13 Glu→Gln human insulin

The self-rotation function of the B13Q mutant (space group R3) shows 60° repeats of two sets of non-crystallographic twofold axes (see figure 3.5). The most prominent set has peaks of 26.4% of the height of the origin peak, at (ω,ϕ)=(90,(12+n.60)) with n=0,1,...,5. The second set has slightly lower height (22.7%) at (ω,ϕ)=(90,(44+n.60)). These twofold axes confirm the approximate 32 symmetry of the hexamer. The hexamer is built up of three exact dimers around the threefold axis of the space group, which is along the z-axis. The dimers' non-crystallographic twofold axes are perpendicular to the crystallographic threefold axis, and thus lie in the xy-plane.

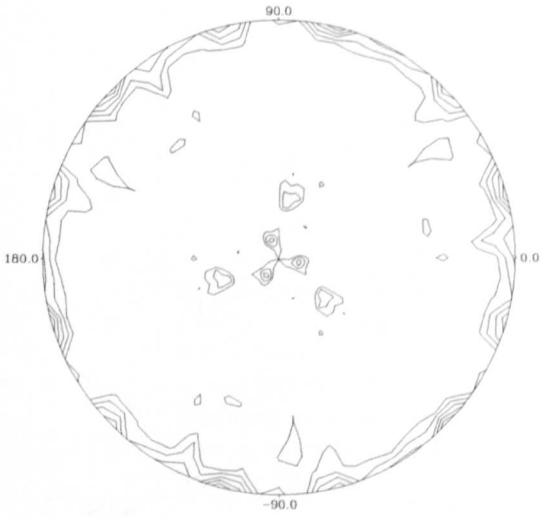

Figure 3.5: Self-rotation function of mutant B13Q; section κ=180°. ω runs from 0 to 90° from the centre to the perimeter, ϕ runs from 0 to 360° around the perimeter.

Zn^{2+}→Co^{2+} human insulin

The the self-rotation functions of the isomorphous CoI and B13Q insulins, both rhombohedral structures in T3R3-conformation, are very similar, as expected. The

only significant differences are the order and peak height of the peaks for the twofold symmetry. In CoI insulin the most prominent set of twofold peaks (see previous section) has a height of 23.5% and ϕ-angles of (44+n.60)° with n=0,1,...,5, while the second set has a height of 22.5% and ϕ-angles of (12+n.60)°.

B9 Ser→His human insulin, rhombohedral form

The B9H mutant in the rhombohedral form is isomorphous with both CoI and B13Q insulin, and therefore has similar self-rotation patterns. The set of twofold peaks with ϕ-angles of (44+n.60)° has a peak height of 35.5%, the set with ϕ=(12+n.60)° has a peak height of 33.5%.

B9 Ser→His human insulin, monoclinic form

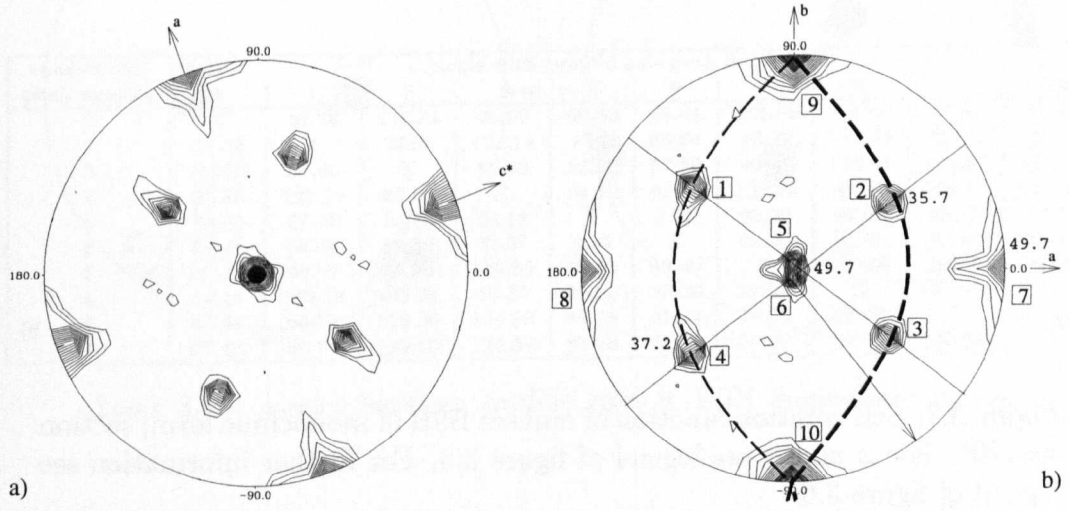

Figure 3.6: Self-rotation function of mutant B9H in monoclinic form; section κ=180°. For ω and ϕ see legend of figure 3.5. Arrow indicates the direction of the threefold axis perpendicular to one set of twofold axes.
a) orthogonalisation such that the monoclinic axis b is along z, perpendicularly out of the plane of the paper, with a and c^* in the plane of the paper, as indicated;
b) orthogonalisation such that c^* is perpendicularly out of the plane of the paper, with a and b along x and y, respectively, as indicated.

The asymmetric unit of the B9H mutant in the monoclinic form contains a full R6 hexamer, for which the non-crystallographic twofold and threefold symmetry is expected to show clearly in a self-rotation. The non-crystallographic twofold symmetry is quite strong (figure 3.6a and b). The highest non-crystallographic peak on section κ=180° has a height of 49.7% of the origin peak, with (ω,ϕ)=(90,110.4) in figure a and (90,0) in figure b. This peak indicates a twofold axis along the crystal a-axis, and shows the monoclinic angle of 110.4°(figure 3.6a). This axis is one of the non-crystallographic dimer axes, or rather its symmetry-equivalent, which is along the reciprocal c-axis. The other two dimer axes are in more general positions in the self-rotation (see figure 3.6b). The heights of the peaks are indicated.

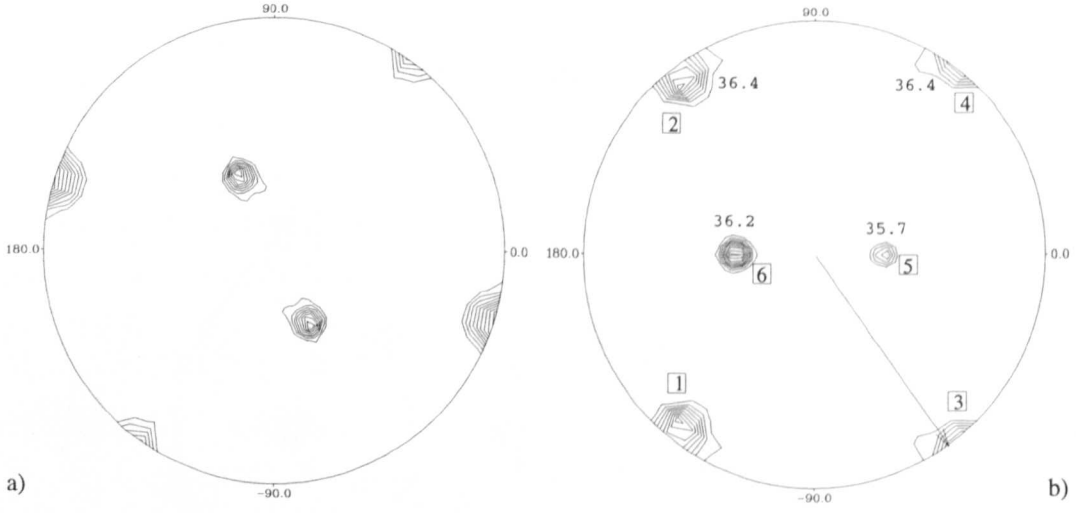

Figure 3.7: Self-rotation function of mutant B9H in monoclinic form; section κ=120°. For ω and ϕ see legend of figure 3.5. For further information see legend of figure 3.6.

The non-crystallographic threefold symmetry of the hexamer is shown in figure 3.7a and b. The full symmetry of the insulin hexamer is 32, as revealed by the combination of three twofold axes roughly perpendicular to a threefold axis in the direction indicated by arrows in figures 3.6b and 3.7b. In the present mutant, like in native monoclinic insulin, this local threefold axis is perpendicular to the reciprocal

c-axis. There is a peak for rotation of dimer 1 onto dimer 2, and one for dimer 1 onto dimer 3, and two more peaks because of the space group symmetry.

Additional threefold symmetry is featuring on the section κ=120°, with additional roughly appropriately spaced twofold axes on section κ=180°at roughly 90°to the threefold axes, indicated by dashed lines in figure 3.6b. The thick dashed line combines with the higher peak (36.2% of the origin peak) on the left on the 0-180 meridian in figure 3.7b, revealing additional 32 symmetry; the actual κ-angle of this peak is 121.4°. The thin dashed line combines with the lower peak (35.7%) on the right; the actual κ-angle of this peak is 110.9°. This 32 symmetry is slightly artificial, since it involves translational symmetry from the space group 2_1 screw axis, and is therefore difficult to confirm with packing diagrams. A dimer from one hexamer will be rotated onto another dimer in another hexamer. The angles between the twofold and threefold axes are shown in tables 3.10 and 3.11.

twofold axis	angle with twofold axis peak number (°)									
peak number	1	2	3	4	5	6	7	8	9	10
1	X	82.78	118.34	95.80	60.33	61.41	141.69	38.15	67.43	67.43
2	82.78	X	58.60	121.14	57.23	58.64	46.22	133.78	60.71	60.71
3	118.34	58.60	X	87.92	58.79	57.39	46.09	133.91	119.30	119.30
4	95.80	121.14	87.92	X	64.12	62.57	133.65	46.35	124.59	124.59
5	60.33	57.23	58.79	64.12	X	2.35	90.00	90.00	88.78	88.78
6	61.41	58.64	57.39	62.57	2.35	X	90.00	90.00	91.22	91.22
7	141.69	46.22	46.09	133.65	90.00	90.00	X	180.00	90.00	90.00
8	38.15	133.78	133.91	46.35	90.00	90.00	180.00	X	90.00	90.00
9	67.43	60.71	119.30	124.59	88.78	91.22	90.00	90.00	X	180.00
10	67.43	60.71	119.30	124.59	88.78	91.22	90.00	90.00	180.00	X

Table 3.10: Angles between twofold axes in B9H monoclinic insulin

B8 Gly→Ser, B13 Glu→Gln, B30 Thr-NH_2 human insulin

Since the B9H and B8S/B13Q/B30amide mutants in the monoclinic crystal form are isomorphous, the self-rotation functions are very similar. Section κ=180°shows the monoclinic angle of 110.5, sufficiently different from the monoclinic angle of B9H insulin to be distinguishable in the peak list of the self-rotation function. The

threefold axis peak number → twofold axis peak number ↓	1	2	3	4	5	6
1	78.08	-	-	102.33	91.68	-
2	-	90.24	89.66	-	-	89.83
3	90.40	-	-	89.70	-	90.00
4	-	89.87	90.09	-	90.51	-
5	87.55	-	-	92.60	-	-
6	-	87.40	92.45	-	-	-
9	-	-	-	-	90.00	90.00

Table 3.11: Angles between twofold axes and threefold axes in B9H monoclinic insulin. Only combinations of 32 symmetry tabulated.

peak height for this non-crystallographic symmetry element is 43.5%. The non-crystallographic threefold axis lies perpendicular to the z-axis and has a peak height of 32.6%.

3.5 Refinement

3.5.1 B13 Glu→Gln human insulin

Refinement of B13Q started with the calculation of $(2F_{obs} - F_{calc})$ and $(F_{obs} - F_{calc})$ maps with the magnitudes from the mutant and the phases from the 4Zn native rhombohedral insulin structure. When more highly refined coordinates became available for the native structure, the refinement of the mutant was re-initiated, using some of the knowledge acquired in the previous attempt. An overview of the refinement of B13Q is shown in figure 3.8.

The preliminary refinement attempt indicated that the zinc ion on the axis in molecule 2 is absent and should be removed at the beginning of refinement. Residue B13 was changed from Glu to Gln. The first refinement, starting from an R-factor of 0.264, consisted of an X-PLOR protocol with full heat stage, and resetting of temperature factors for the main chain between 10 and 25Å^2, and for the side chains

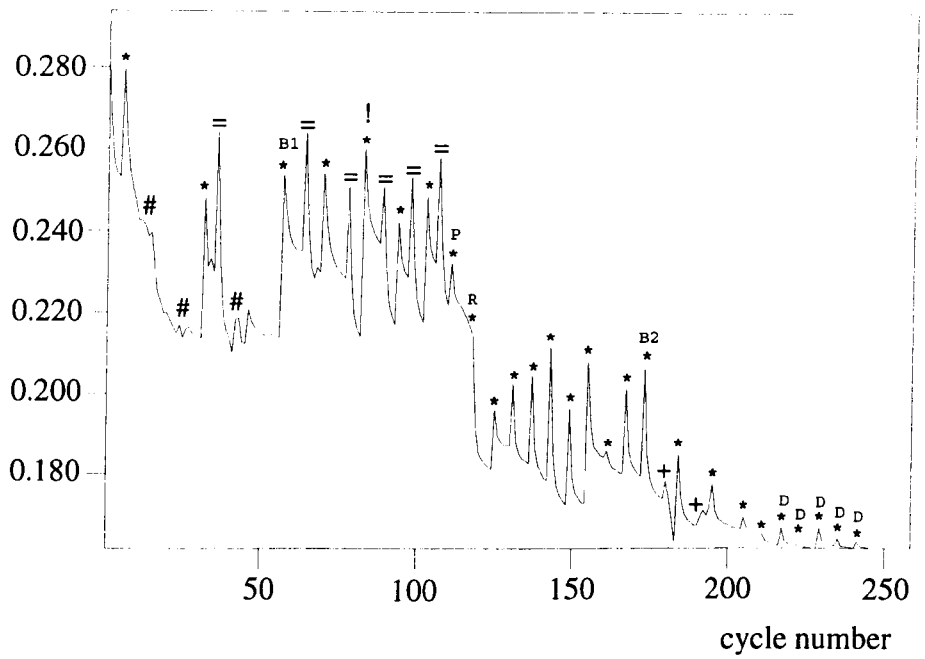

Figure 3.8: Overview of R-factor during refinement of B13Q insulin.
*: manual rebuild, including water; then PROLSQ refinement of coordinates when the R-factor is >0.25, refinement of coordinates and temperature factors at lower R-factors;
#: adapting the matrix value for X-ray relative to geometry contribution between cycles of PROLSQ refinement;
=: resetting temperature factors prior to PROLSQ refinement (see *);
B1: first indication of the need for swapping the positions of B1/B2 main chain (molecule 2) and B3 side chain;
B2: B1-B3 (molecule 2) satisfactorily refined to new positions;
!: changing the appropriate waters to chloride ligands for the zinc ions;
P: PROCHECK used to aid rebuilding for the first time;
R: changing from using resolution limits 10-1.9Å to 10-2.3Å in PROLSQ refinement;
+: ARP refinement of water structure, incorporating PROLSQ;
D: investigating disorder at B13 (molecule 1).

between 10 and 30Å^2. The maps then confirmed the absence of the axial zinc in molecule 2. Spikes upwards in figure 3.8 are commonly caused by manual rebuilding of the structure with FRODO, after which least-squares refinement was performed using PROLSQ. In the initial stages, at R-factors higher than around 0.25, only atomic coordinates were refined, while at lower R-factors temperature factors were allowed to refine isotropically. Temperature factors were reset at various stages to free the refinement from local minima. The restraints on geometry were adjusted by adapting the relative weight of the contribution of the X-ray data.

At stage B1 (see figure 3.8) it became clear that the N-terminus of the B chain of molecule 2 was poorly defined, after which many omit maps were calculated (by setting the occupancy of appropriate atoms to zero and performing a few cycles of refinement with the reciprocal space refinement program used at the time) and rebuilding was attempted. The introduction of information from PROCHECK during the rebuilding stages made the spikes upwards smaller on the whole, indicating more geometrically acceptable changes so that less refinement of the geometry was required by PROLSQ. The highest resolution data were not as good and complete as originally assumed, leading to the change in resolution limits at point R. Refinement of the water structure was made easier and more successful when the automatic refinement program ARP (Lamzin and Wilson, 1992) became available. Refinement of the disorder at B13 (molecule 1) was approached in separate concurrent refinement runs: in run A one conformation of the side chain was included with water in the remaining electron density, in run B *vice versa.*

The final R-factor for all data between 25.65 and 2.3Å resolution is 0.182. For restraint and geometry information, see table 3.13 at the end of this section.

3.5.2 $Zn^{2+}\rightarrow Co^{2+}$ human insulin

The starting model for refinement of CoI insulin was the native rhombohedral zinc insulin molecule in T3R3-conformation. Preliminary investigations of the modified

molecule indicated the presence of two metal ions on the threefold rhombohedral axis, and a phenol molecule in the position resembling that found in monoclinic structures in the R6-state (Derewenda *et al.*, 1989), which in the native rhombohedral T3R3-structure is roughly where the off-axial zinc is situated. An overview of the refinement procedure is shown in figure 3.9.

The first step consisted of rigid body refinement of the protein component of the molecule in X-PLOR, where the dimer model was refined as a whole rigid body first, after which it was split up into monomers. Figure 3.9 starts with the R-factors for the first reciprocal space refinement cycle, in which the metal ions (zinc, in the initial stages) and phenol molecule were included. The R-factor for all data to 1.9Å was 0.2982 at this stage. The C-termini of both B chains and the N-terminus of the B chain of molecule 2 were excluded from refinement because of the obvious need for rebuilding.

At point C (see figure 3.9) the metal ions were changed from zinc to cobalt, and the axial ligand of the metal ion in molecule 2 was changed from water to iodide. PROLSQ refinement followed, first without keeping a special distance between the metals and ligands, and then a few cycles with a special distance imposed between cobalt and iodide. The distances between cobalt and histidine were allowed to refine freely.

At point S the density for the N-terminus of the B chain of molecule 2 became interpretable. The side chain of residue B3 was swapped with the main chain of residues B1/B2, changing the last turn of the α-helix to an extended conformation. At point L a second phenol molecule was included, which has hydrogen bonds with B10 His $N\delta_1$ and B13 Glu $O\epsilon_1$. The total number of phenol molecules in this T3R3 hexamer is therefore six. From point R the data between 1.9 and 1.95Å were excluded from refinement because of problems caused by the incompleteness of that resolution shell of data.

At point G all but the three most obvious waters, one of which is liganded to

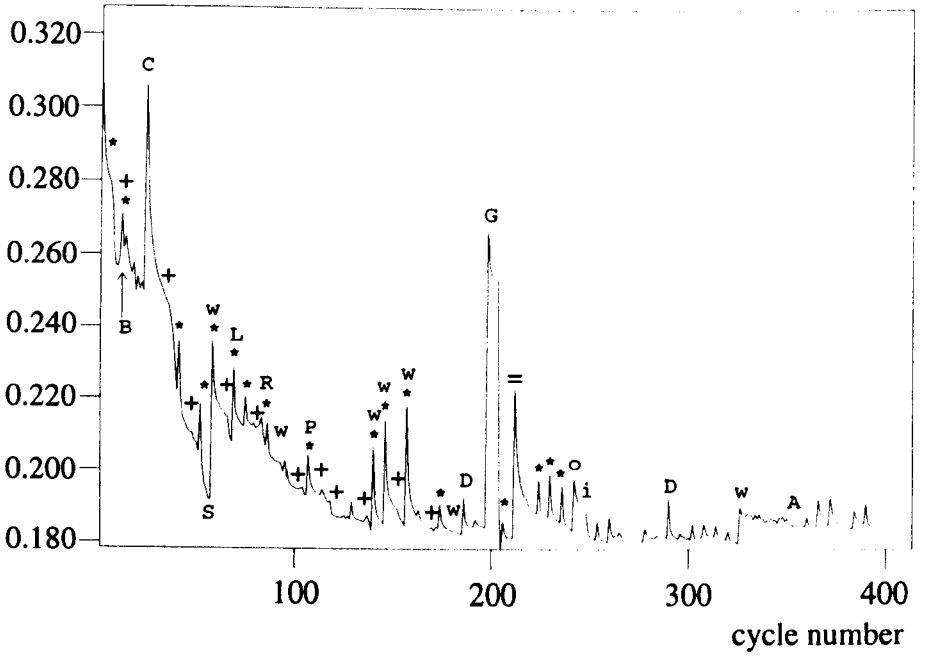

Figure 3.9: Overview of R-factor during refinement of CoI insulin.
*: manual rebuild, including water; then PROLSQ refinement of coordinates only in the first stages, refinement of both coordinates and temperature factors after point B;
B: changing from coordinates-only to coordinates-and-temperature-factor refinement;
+: ARP refinement, incorporating PROLSQ.
S: swapping the side chain of B3 with the main chain of B1/B2;
w: revising the water structure;
L: second phenol molecule included with hydrogen bonds to B10 and B13;
P: PROCHECK information aiding rebuilding for the first time;
D: sorting out disorder of residue D21 (molecule 1);
=: resetting temperature factors prior to PROLSQ refinement;
o: residues B1-B3 omitted;
i: residues B1-B3 included again;
A: probing of the density around position (0,0,15) (orthogonal Å) with various cations and anions present in the crystallisation solution.

the cobalt ion in molecule 1, were deleted, after which the water structure was rebuilt in five runs of ARP. R-factor data were only recorded for the last run: down from 0.1920 to 0.1783. Refinement cycle 250 to 350 were taken up with refinement of the water structure by detailed rebuilding and reciprocal space refinement. From point A onwards it was tried to model a threefold degenerate acetate ion on the threefold axis on the surface of the hexamer, in the shallow depression which is lined by the three B chain helices in the R-conformation half of the hexamer. The density around (0,0,15) (orthogonal Å) was probed with cobalt and sodium cations, and acetate, citrate, iodide and chloride anions. None of the combinations of cat- and anions gave a totally satisfactory result judged by the electron density after refinement, also reflected by very similar R-factors for every conformation investigated. The density most resembles an acetate ion, probably coordinated to sodium. The final R-factor for CoI insulin is 0.1933 for all data between 33.33 and 1.95Å. For restraint and geometry information, see table 3.13 at the end of this section.

3.5.3 B8 Gly→Ser, B13 Glu→Gln, B30 Thr-NH_2 human insulin

With the advent of new techniques and ideas in protein crystallography, and better model coordinates being made available, it seemed necessary to re-initiate refinement of the B8S/B13Q/B30amide a few times. Some of the experience gained in preliminary stages, was used explicitly at the beginning of the final refinement procedure.

The hexameric starting model for the refinement was built up from three native monoclinic dimers from a molecular replacement search with *AMoRe*. The packing within the hexamer should resemble the mutant hexamer more closely than the original hexameric native monoclinic model. The R-factor for this model was 0.4014, while the free R-factor for the 10% of data to be excluded from the refinement, was 0.3953. An overview of the full refinement procedure is shown in figure 3.10.

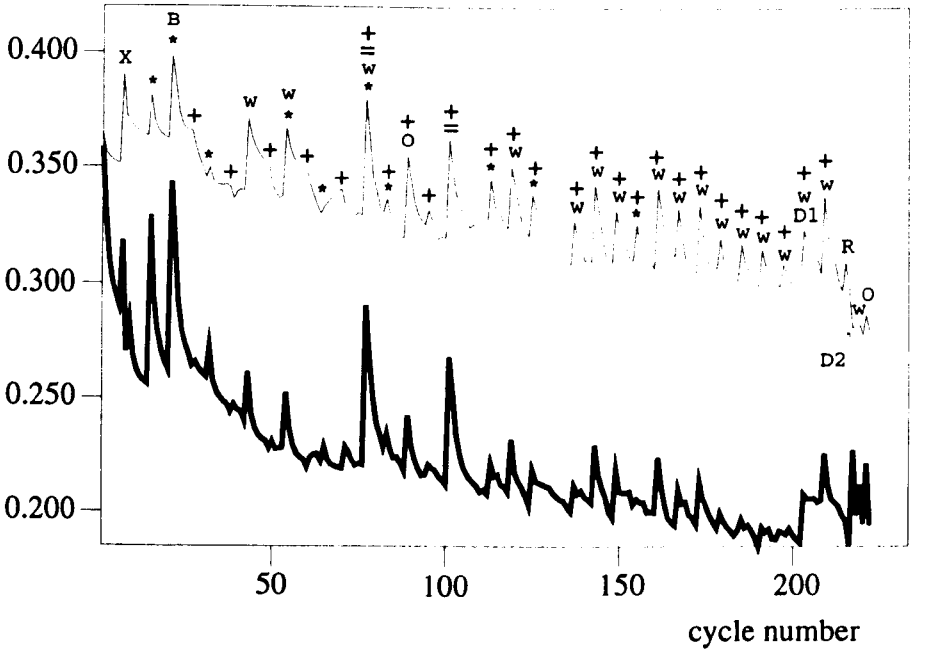

Figure 3.10: Overview of R-factors during refinement of B8S/B13Q/B30amide insulin, conventional R-factor in thick line, free R-factor in thin line.
*: manual rebuild, including water; then, unless otherwise indicated, PROLSQ refinement of coordinates only in the first stages, refinement of both coordinates and temperature factors after point B;
=: resetting temperature factors prior to PROLSQ refinement (see *);
B: changing from coordinates-only refinement to coordinates-and-temperature factor refinement;
X: X-PLOR refinement followed by PROLSQ coordinate refinement;
+: ARP refinement, incorporating PROLSQ. Water molecules suggested by ARP were rejected;
w: developing the water structure with X-SOLVATE, followed by reciprocal space refinement with PROLSQ;
O: rebuilding on the basis of overlaps of all the A and B chains on each other to keep check of the pseudo-symmetry;
D1: deleting all waters with temperature factor higher than 70Å^2;
D2: deleting all waters with temperature factor higher than 90Å^2;
R: reciprocal space refinement of coordinates and temperature factors performed with REFMAC for the first time. Four cycles of refinement in every stage, only first and last R-factors indicated.

The first step in the refinement consisted of real space refinement of all the main chain peptide bonds and all the side chains with X-AUTOFIT (Oldfield, 1996). Previously determined coordinates for the zinc ions, chlorides and phenol molecules were included at this point. The C-terminal residues of all six B chains, the N-terminal residues of chains H and L, the atoms of E1 and I1 and the side chains of residues A14, D1 and I14 were excluded from the subsequent first round of reciprocal space refinement with PROLSQ. Refinement stage X (see figure 3.10) was an extensive X-PLOR protocol, with full occupancy for all available atoms. The protocol consisted of rigid body refinement of the hexamer as a whole, then of the dimers separately, followed by the monomers. The subsequent simulated annealing run was of the 'slowcool' type, without full equilibration at high temperatures. This was immediately followed by PROLSQ coordinate refinement. Stages of manual rebuilding were aided by information from PROCHECK. The water structure was built up with X-SOLVATE in QUANTA. After point R (see figure 3.10) reciprocal space refinement was performed with REFMAC, using the maximum likehood residual for structure factors. The R-factor for the last point in figure 3.10 is 0.194 (free R-factor 0.280), calculated for the data between 11.0 and 2.1Å. The final R-factor for all data between 43.44 and 2.1Å is 0.215, the free R-factor 0.306. For restraint and geometry information, see table 3.14 at the end of this section.

3.5.4 B9 Ser→His human insulin

Monoclinic crystal form

The starting model for refinement of B9H insulin in monoclinic form was constructed from three native monoclinic dimers, positioned in the mutant unit cell with *AMoRe*. This model should resemble the mutant structure more closely than the native hexamer model positioned in the mutant unit cell. The R-factor of this model (which contains no ligands of any kind) is 0.4007, while the R-factor for the

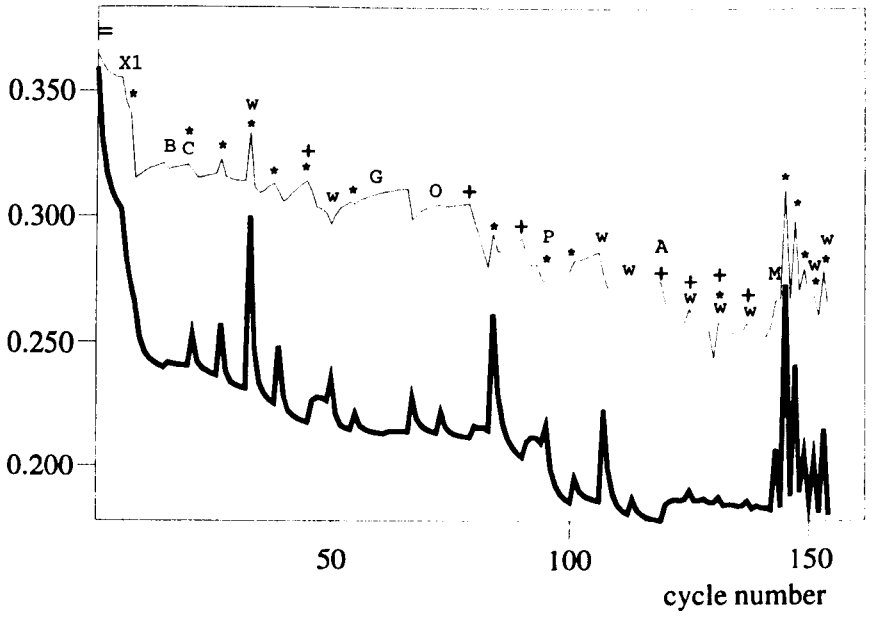

Figure 3.11: Overview of R-factors during refinement of B9H insulin, monoclinic form, conventional R-factor in thick lines, free R-factor in thin lines.
*: manual rebuild, including water; then, unless otherwise indicated, PROLSQ refinement of coordinates only in the first stages, refinement of both coordinates and temperature factors after point B;
=: resetting temperature factors prior to PROLSQ refinement (see *);
B: changing from coordinates-only refinement to coordinates-and-temperature factor refinement;
X1: X-PLOR minimisation (coordinates and temperature factors), no molecular dynamics involved;
C: changing the fourth ligands for both zinc ions to chloride;
+: ARP refinement, incorporating PROLSQ;
A: ARP refinement, after which the water molecules suggested are removed from the structure;
w: developing the water structure with X-SOLVATE, followed by reciprocal space refinement with PROLSQ;
G: reducing the restraints on the geometry slowly over 35 cycles of refinement;
O: rebuilding on the basis of overlaps of all the A and B chains on each other to keep check of the pseudo-symmetry;
P: using PROCHECK to aid manual rebuilding for the first time;
M: reciprocal space refinement of coordinates and temperature factors performed with REFMAC for the first time. Four cycles of refinement in every stage, only first and last R-factors indicated.

10% of reflections to be excluded from refinement, is 0.4424. An overview of the refinement procedure is given in figure 3.11.

The first maps gave clear indications for the positions of the B9 mutations, the presence of two zinc ions and their fourth ligands on the non-crystallographic three-fold axis, the presence of phenol molecules very similar to the native structure, and some of the terminal residues that had been omitted from the starting model. This information was included in the first reciprocal space refinement, the starting point of figure 3.11. In the first stages, only coordinates were refined. Point X1 represents two cycles of minimisation in X-PLOR. The water structure was investigated both with ARP and X-SOLVATE. After point A ARP was only used for refinement, while any water molecules suggested by ARP were not included in the structure.

From point M reciprocal space refinement was performed with REFMAC, using the maximum likelihood residual for structure factors. At that stage it became clear that some of the electron density peaks in the core of the hexamer represented zinc ions rather than water molecules. The R-factors of the final model are 0.2156 (conventional R-factor) and 0.3167 (free R-factor), for all data between 57.74 and 2.1Å. The refinement was performed with all data between 13.0 and 2.1Å. For restraint and geometry information, see table 3.14 at the end of this section.

Rhombohedral crystal form

B9H rhombohedral insulin is isomorphous with native 4Zn insulin. A starting model for refinement was derived from a molecular replacement search with *AMoRe*, which confirmed isomorphism. The advent of new techniques and ideas has had a great impact on the process of refinement of the B9H mutant. Refinement of the mutant against data collected at room temperature was not followed through when cryocrystallographic techniques were introduced. However, the success of the first attempt at collecting data at 120K was limited. The crystal was not very stable in the cryoprotectant solution, and upon freezing only diffracted to around 2.5Å. These data

were complete and reliable to 2.6Å resolution. A later attempt at data collection was more successful, when more experience had been gained with chosing a cryoprotectant and the techniques of flash freezing. Crystals were submerged in a solution with 1,2-ethanediol instead of glycerol which had been the earlier choice, and one of these diffracted well to 1.9Å resolution. Data sets could not be merged because of non-isomorphism due to temperature effects upon freezing (see table 3.12).

Refinement of the mutant with REFMAC was started with a model including zinc ions on the rhombohedral axis, and indications for chloride ions and other zinc ions in the form of water molecules. Unfortunately, REFMAC is as yet incapable of dealing effectively with restraining special distances when atoms in special positions and atoms from symmetry-related molecules are involved. The refinement of the rhombohedral B9H mutant has therefore stagnated at R-factors of 0.2338 (for the working set, data between 17.08 and 1.92Å) and 0.3097 (for the 10% of reflections kept 'free'). Refinement with X-PLOR was only effective in the earliest stages of refinement, not for the detailed refinement of water structure and the intricate zinc ion structure of the mutant. For restraint and geometry information, see table 3.13 at the end of this section.

data set characteristics			unit cell edges	
temperature	resolution (Å)	effective resolution (Å)	a=b= (Å)	c= (Å)
room temp	2.0	2.3	80.79	37.47
120K	2.5	2.6	79.84	36.65
120K	1.9	1.92	79.878	37.427

Table 3.12: Unit cell parameter change upon freezing B9H rhombohedral insulin

3.5.5 B9 Ser→Asp, B27 Thr→Glu human insulin

The starting model for B9D/B27E insulin was produced with *AMoRe*, although the mutant is isomorphous with native orthorhombic insulin. The refinement is

described below, with an overview shown in figure 3.12.

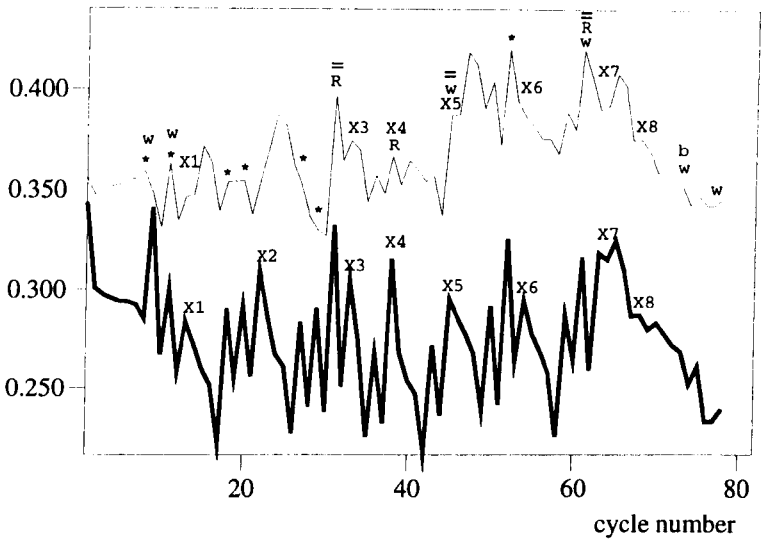

Figure 3.12: Overview of R-factors during refinement of B9D/B27E insulin, conventional R-factor in thick lines, free R-factor in thin lines.
*: manual rebuild, including water; then, unless otherwise indicated, REFMAC refinement of coordinates and temperature factors;
w: developing the water structure with X-SOLVATE, followed by reciprocal space refinement with REFMAC;
=: resetting temperature factors prior to REFMAC refinement (see *);
X1,X3,X4,X5,X6,X7: X-PLOR full refinement protocol (slowcool), immediately followed by REFMAC;
X2: X-PLOR full refinement protocol (slowcool) with all non-water atoms included with occupancy 1.0;
X8: X-PLOR full refinement protocol (slowcool), followed by 4 cycles of REFMAC; no waters included, experimental standard deviations not used;
b: temperature factors of water molecules reset to average temperature factor of protein.

The conventional crystallographic R-factor for the starting model was 0.5119 for all data to 2.3Å. Ten percent of the reflections were excluded from the refinement procedure, reserved for calculations of the free R-factor. Reciprocal space refinement was performed with REFMAC and X-PLOR. In the second cycle of X-PLOR refinement all atoms present in the coordinate set (apart from water molecules) were

refined, without omitting uncertain ones. After X-PLOR step 8 it was discovered that the low resolution reflections produced erratic statistics. Some trials were done where experimental standard deviations were not used in the scaling of reflections, different resolution limits were used, and ultimately the idea of using the free R-factor was given up. The refinement of 829 atoms (at the last point of figure 3.12) with 3632 reflections, keeping 384 reflections separate for free R-factor calculations, is not warranted. It seemed justified to use the test reflections for scaling purposes in REFMAC until the number of parameters became too large.

The unpredictable behaviour of both R-factors is probably due to the fact that the structure factors calculated from the model are all systematically lower (in the range used for refinement) than the observed structure factors (see figure 3.13), which is caused by the significant differences of just a few low resolution reflections. This must be due to incorrect measurements or a systematic error.

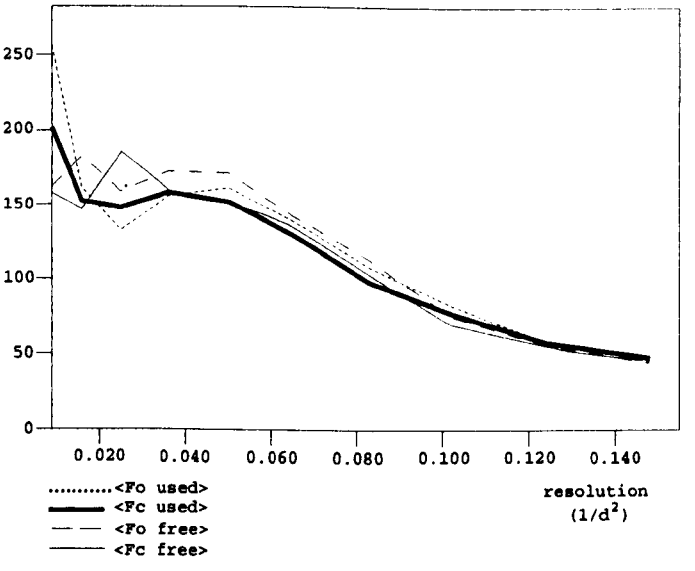

Figure 3.13: Indication of problems with a small subset of reflections. In the resolution range used for refinement, all reflections have F_{obs} higher than F_{calc}, which can not be corrected with a scale factor because of the low resolution reflections. Structure factors on arbitrary scale.

The conventional R-factor at the end of the refinement procedure for B9D/B27E insulin for all data between 13.0 and 2.3Å is 0.2286. The R-factor is 0.3265 for all data between 29 and 2.3Å, further illustrating the problem at low resolution. For restraint and geometry information, see table 3.14 at the end of this section.

3.5.6 A21 Asn→Gly, B9 Ser→Glu, B10 His→Glu human insulin

The molecular replacement study of A21G/B9E/B10E insulin revealed a new packing of insulin dimers. The data collected at room temperature extended to 2.8Å resolution, which made detailed refinement impossible. The data collected at 120K were incomplete and of insufficient quality. These data could not be merged with the data collected at room temperature because of non-isomorphism.

Although the results of refinement of a protein structure at 2.8Å resolution have to be interpreted with extreme care, the process can usually be done very rapidly. An attempt was made with the A21G/B9E/B10E mutant, when the maps after molecular replacement, and rigid body refinement and minimisation in X-PLOR proved easily interpretable. The mutant aggregates in dimer form very closely resembling that of the native orthorhombic molecule, but in different packing. This puts the C-termini of the B chains in close contact with symmetry-related molecules. The maps, however, clearly showed the possibilities for change. Also, the mutations could be modelled in these maps. The first stage of reciprocal space refinement brought the R-factors down to 0.286 and 0.370. After rebuilding the protein molecules and modelling some of the solvent structure, the 10% 'free' reflections were included in the refinement to increase the number of reflections for every parameter refined. Within four steps of six minicycles of REFMAC refinement, the R-factor had dropped to 0.182. An overview of the limited refinement of the mutant is shown in figure 3.14. For restraint and geometry information, see table 3.14 at the end of this section.

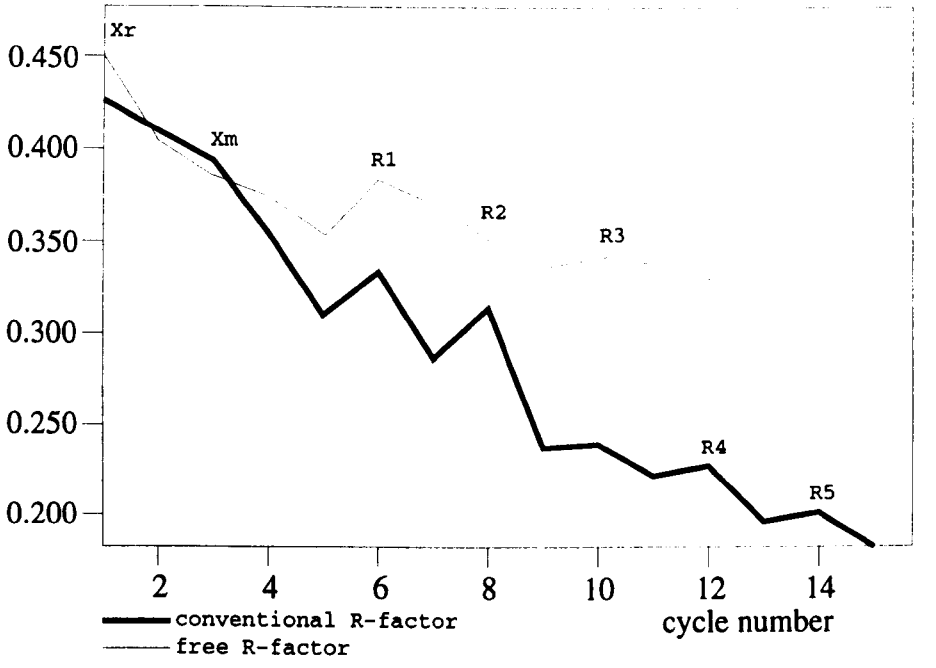

Figure 3.14: Overview of R-factors during refinement of A21G/B9E/B10E insulin.
Xr: X-PLOR rigid body refinement of dimers, monomers, and B chain C-termini;
Xm: X-PLOR minimisation (including temperature factors);
R1: REFMAC refinement of coordinates and temperature factors, followed by manual rebuilding;
R2: REFMAC refinement of coordinates and temperature factors, preceded by resetting of temperature factors;
R3: REFMAC refinement of coordinates and temperature factors, preceded by manual rebuilding of protein and developing water structure with X-SOLVATE;
R4: REFMAC refinement of coordinates and temperature factors using all observed reflections, preceded by manual rebuilding of protein and developing water structure with X-SOLVATE;
R5: REFMAC refinement of coordinates and temperature factors.

3.5.7 A4 Glu→Gln, B25 Phe→Tyr, des B30 single chain human insulin

The B25Y(B29-A1)A4Q mutant crystallised in a space group for which no detailed structure has been published. Unfortunately the mutant did not diffract well enough to warrant detailed refinement. The R-factor of the molecular replacement model was 0.5767 for all data to 2.8Å. Since there were only 1386 reflections, no free R-factor was used in the refinement.

The mutations from Phe to Tyr and from Glu to Gln could not be interpreted unambiguously at this resolution. However, the modifications at the C-terminus of the B chain were clearly defined in maps. Residue B30 was deleted, after which a peptide bond was created between residues B29 and A1, the density for which was clear and continuous. Also, some strong $(F_{obs} - F_{calc})$ density was interpretable as a zinc ion close to B10 histidine, with the space group symmetry such that three histidines coordinate to this zinc ion. The coordinates were refined with X-PLOR and PROLSQ, but refinement was unstable and unreliable at 2.8Å resolution. The refinement was stopped at an R-factor of 0.3728 for all observed data, which lie between 11.6 and 2.8Å.

rms deviations	B13Q	B9H rhombohedral	CoI
DISTANCE RESTRAINTS			
bond length	0.014 (0.020)	0.024 (0.020)	0.025 (0.020)
angle distance	0.045 (0.040)	0.041 (0.040)	0.050 (0.020)
planar distance	0.052 (0.050)	0.043 (0.050)	0.052 (0.050)
PLANAR GROUPS			
peptide plane	0.0267 (0.0300)	0.0253 (0.0300)	0.0290 (0.0300)
aromatic plane	0.0148 (0.0200)	0.0154 (0.0200)	0.0162 (0.0200)
CHIRAL VOLUME	0.145 (0.150)	0.171 (0.150)	0.197 (0.150)
NON-BONDED CONTACTS			
single torsion	0.200 (0.300)	0.185 (0.300)	0.187 (0.300)
multiple torsion	0.278 (0.300)	0.254 (0.300)	0.272 (0.300)
Hbond	-	0.226 (0.300)	0.316 (0.300)
TORSION ANGLES			
planar	6.6 (7.0)	4.8 (7.0)	5.2 (7.0)
staggered	18.9 (15.0)	21.0 (15.0)	19.2 (15.0)
transverse	28.7 (20.0)	20.5 (20.0)	16.1 (20.0)
THERMAL FACTORS			
main chain bond	2.217 (2.000)	2.661 (2.000)	2.547 (2.000)
main chain angle	3.452 (2.500)	4.065 (3.000)	3.624 (2.500)
side chain bond	3.715 (3.000)	2.438 (2.000)	4.229 (3.000)
side chain angle	5.520 (4.500)	3.838 (3.000)	5.671 (4.500)

Table 3.13: Geometry restraints of the rhombohedral modified insulin structures. Root mean square deviations, with target values in brackets.

rms deviations	B8S/B13Q/B30amide	B9H monoclinic	B9D/B27E	A21G/B9E/B10E
DISTANCE RESTRAINTS				
bond length	0.019 (0.020)	0.022 (0.020)	0.019 (0.020)	0.023 (0.020)
angle distance	0.042 (0.040)	0.045 (0.040)	0.053 (0.040)	0.058 (0.040)
planar distance	0.048 (0.050)	0.050 (0.050)	0.051 (0.050)	0.059 (0.050)
PLANAR GROUPS				
peptide plane	0.0244 (0.0300)	0.0284 (0.0300)	0.0286 (0.0300)	0.0324 (0.0300)
aromatic plane	0.0140 (0.0200)	0.0158 (0.0200)	0.0121 (0.0200)	0.0146 (0.0200)
CHIRAL VOLUME	0.184 (0.150)	0.177 (0.150)	0.167 (0.150)	0.181 (0.150)
NON-BONDED CONTACTS				
single torsion	0.185 (0.300)	0.183 (0.300)	0.216 (0.300)	0.235 (0.300)
multiple torsion	0.262 (0.300)	0.267 (0.300)	0.282 (0.300)	0.271 (0.300)
Hbond	0.244 (0.300)	0.272 (0.300)	0.267 (0.300)	0.213 (0.300)
TORSION ANGLES				
planar	4.2 (7.0)	4.8 (7.0)	4.1 (7.0)	5.7 (7.0)
staggered	21.6 (15.0)	20.0 (15.0)	22.5 (15.0)	28.9 (15.0)
transverse	24.1 (20.0)	19.7 (20.0)	34.3 (20.0)	27.2 (20.0)
THERMAL FACTORS				
main chain bond	3.142 (2.000)	3.251 (2.000)	3.609 (2.000)	4.970 (2.000)
main chain angle	5.235 (3.000)	4.688 (3.000)	5.270 (3.000)	7.141 (3.000)
side chain bond	2.891 (2.000)	3.406 (2.000)	5.380 (3.000)	4.550 (2.000)
side chain angle	4.503 (3.000)	5.340 (3.000)	7.244 (4.000)	6.793 (3.000)

Table 3.14: Geometry restraints of four non-rhombohedral structures. Root mean square deviations, with target values in brackets.

3.6 Description of the structure models

The main features of the three-dimensional structure of insulin are seen in all of the modified insulins described in this thesis. The insulin monomer is held together by three disulfide linkages at A6-A11, A7-B7 and A20-B19. The α-helix running from residues B9 to B19 is largely undisturbed by modifications in its sequence. Where aggregation takes place, the dimer interface contains the small antiparallel β-sheet structure of residues B23-B28, and the hexameric structures have metal ions in the cores. N- and C-termini are often disordered, as was already described by Adams *et al.* (1969). The main differences of the modified insulin structures will be described below.

3.6.1 B13 Glu→Gln human insulin

The mutation of Glu to Gln at position B13 is expected to produce a more stable hexamer because of the reduction in charge repulsion in the core, possibly with an elaborate hydrogen bonding 'ring' structure of the six glutamine residues as described by Markussen *et al.* (1987). The mutant readily crystallised in rhombohedral form (space group R3) in T3R3-conformation. The positions of the mutated residues were not completely clear immediately at the beginning of refinement (see figure 3.15), so the sites were excluded from initial refinement. An omit map calculated (as described in section 3.5.1) after completion of refinement confirms the refined positions (see figure 3.16). Residue B13 (molecule 1) is disordered, and the simple twelve-membered 'ring' structure of the six glutamines is not observed. Moreover, the two glutamines in a dimer are within hydrogen bonding distance of each other, but not of neighbouring symmetry-related B13 glutamines. Contacts in the core of the hexamer are mediated through water molecules.

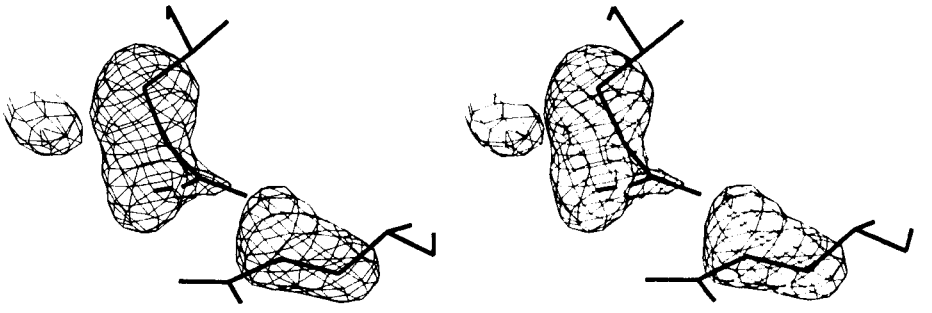

Figure 3.15: Indications for mutated residues in B13Q insulin at the start of refinement. Stereo view of both B13 residues and $F_{obs}-F_{calc}$ density at $+3\sigma$.

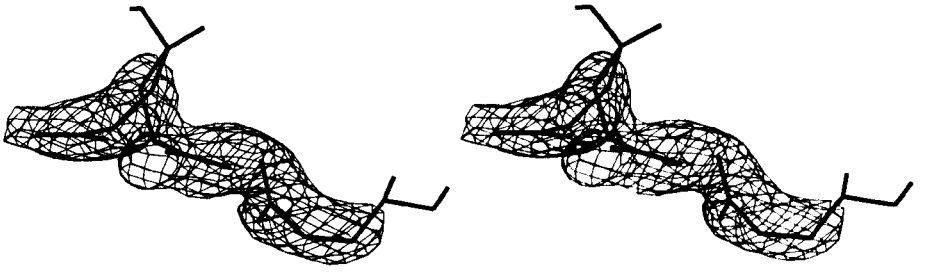

Figure 3.16: Omit maps for mutated residues in B13Q insulin in fully refined positions. Stereo view of both B13 residues and $F_{obs}-F_{calc}$ density at $+3\sigma$, in the same orientation as figure 3.15.

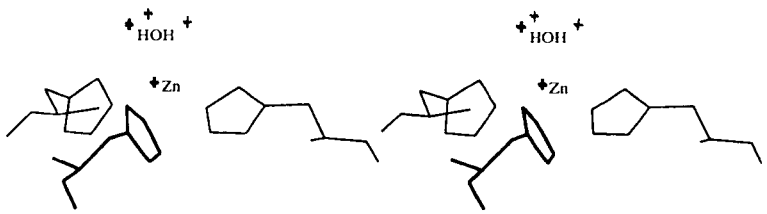

Figure 3.17: Stereo view of zinc in molecule 1 of B13Q insulin, octahedrally coordinated on the threefold axis.

The zinc binding of the mutant only became clear after repeated trials with all zinc ions omitted and various combinations of zinc ions included in early refinement and map calculations. The disorder of the B10 histidine (molecule 2) as seen in native 4Zn insulin, is not present in the B13Q mutant. Instead, the histidine is solely pointing towards the off-axial zinc site. The B13Q mutant is therefore a true 'four zinc' insulin, in contrast with the native structure which actually contains five metal binding sites. The binding of zinc in the axial site in molecule 1, octahedrally coordinated by three symmetry-related B10 histidines and three symmetry-related water molecules, is shown in figure 3.17. The tetrahedral coordination of zinc by B10 histidine, a symmetry-related B5 histidine and two chloride ions is shown in figure 3.18.

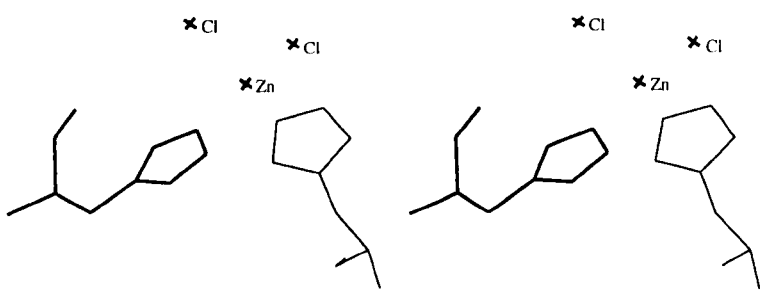

Figure 3.18: Stereo view of zinc in molecule 2 of B13Q insulin, tetrahedrally coordinated.

The termini of the four chains in the asymmetric dimer may be described as follows:

- The A chain N-terminus (molecule 2) is clearly in electron density but has slightly raised temperature factors compared to the average of 33.1Å^2 for the main chain of A. The A chain C-terminus is very clear with temperature factors lower than the average for the residues up to and including A20;

- The B chain N-terminus (molecule 2) has become helical, thus forming an R-conformation, but not completely to residue B1. The side chain of residue B3 takes up the position of what would be the main chain positions of B2 and B1 were the helix to stretch to B1. Residues B1 and B2 exhibit an extended conformation (see figure 4.2). Although the temperature factors of these residues are high at >50 whereas the average of the main chain of B is 30.9Å^2, the electron density gives an indication for the validity of the extended conformation as compared to the full helical conformation (see figure 3.19).

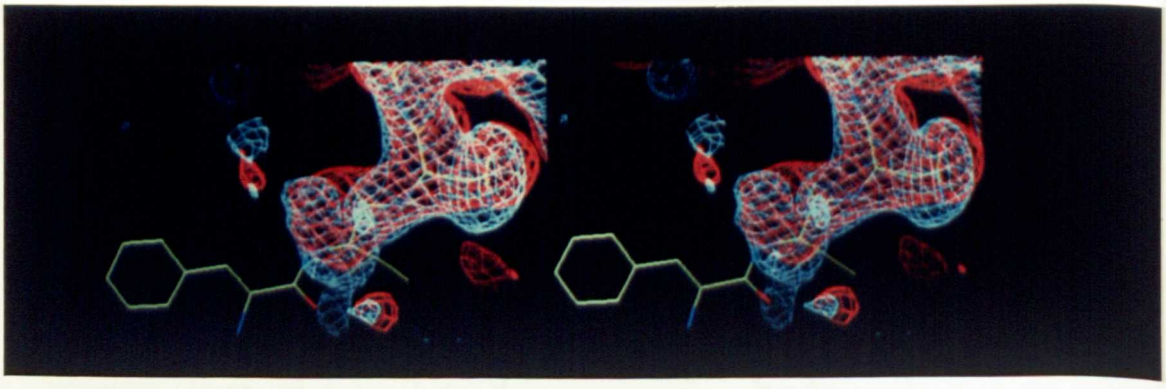

Figure 3.19: Stereo view of the comparison of electron density for two conformations of the N-terminus of molecule 2 in B13Q insulin. $2F_{obs}-F_{calc}$ maps and refined coordinates; orange: map at start of refinement, B1-B3 included in map calculations; light blue: map corresponding to coordinates after rearrangement of the N-terminus to extended conformation, B1-B2 not included in map calculations (coordinates not fully refined).

The Ramachandran plot of the model in the early stages of refinement indicated that residues B2 and B3 were in unacceptable regions (see figure 3.20) whereas the Ramachandran angles of these residues in the final model are within the most favoured regions. The B chain C-terminal residues are clear in the electron density, although they have high temperature factors (53.1-73.5Å^2 for the main chains of the last four residues);

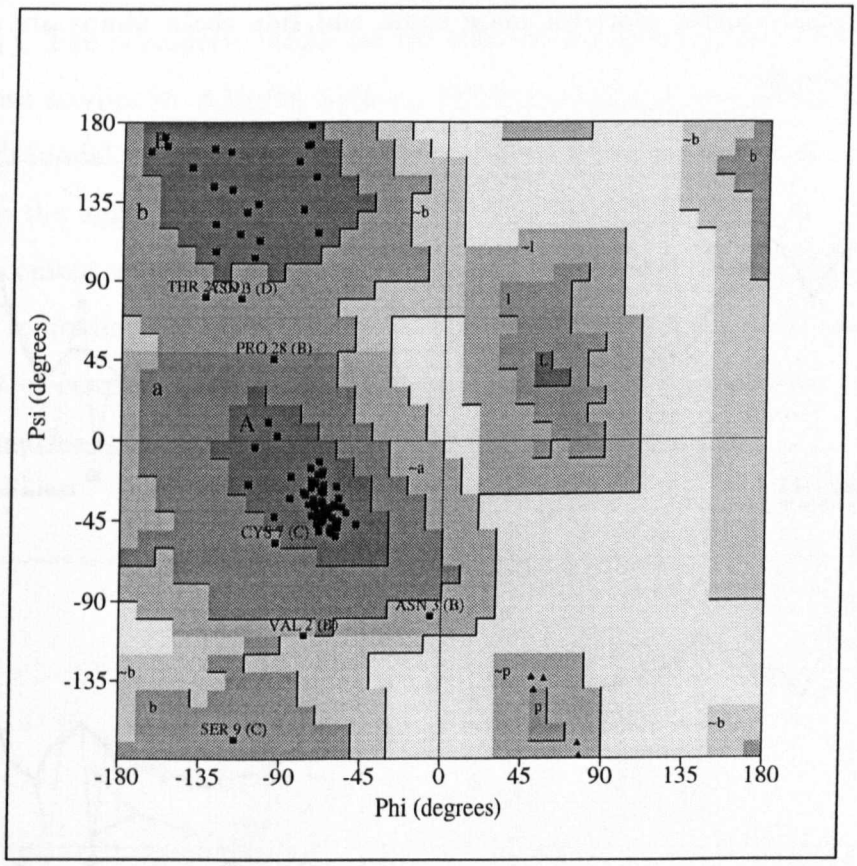

Figure 3.20: Ramachandran plot for the starting model of B13Q insulin. Darkest grey represents the most favoured regions (A,B,L); medium grey represents additional allowed regions (a,b,l,p); lightest grey represents generously allowed regions (~a,~b,~l,~p); white regions are disallowed.

- The C chain termini (molecule 1) have barely elevated temperature factors (average for the main chain of C is 33.0) and are very clear in the electron

density;

- The D chain N-terminus (molecule 1) clearly exhibits an extended T-conformation. The temperature factors of the terminus are elevated gradually towards the end of the chain to a value of 37.5Å^2 for the main chain. The C-terminus has high temperature factors, rapidly increasing to 63.6Å^2 for the main chain of the last residue, but is clear in electron density maps.

Temperature factor plots for main chain and side chain atoms are shown in figures 3.21 a to d.

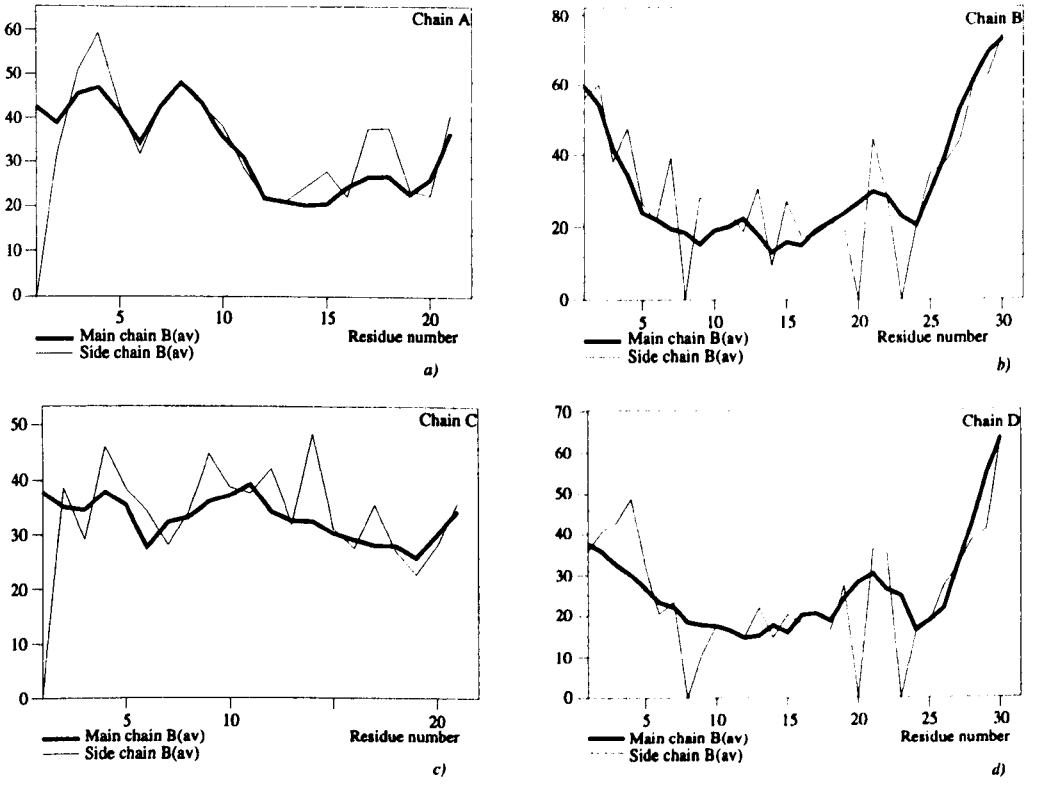

Figure 3.21: Temperature factor plots for the B13Q mutant.
a) Chain A. Average for main chain is 33.1Å^2, for side chain 34.1Å^2.
b) Chain B. Average for main chain is 30.9Å^2, for side chain 34.9Å^2.
c) Chain C. Average for main chain is 33.0Å^2, for side chain 35.2Å^2.
d) Chain D. Average for main chain is 26.2Å^2, for side chain 27.7Å^2.

3.6.2 B9 Ser→His human insulin

Monoclinic crystal form

The mutation of serine to histidine in position B9 allows for extra zinc binding in the core of the insulin hexamer, which it was anticipated should result in increased stability. In order to study zinc binding in the case of all-R-conformation B chains, the B9H mutant was crystallised in monoclinic form, the most stable hexamer (space group $P2_1$). The hexameric model for the mutant was put together from molecular replacement studies in *AMoRe*, with an R2 dimer from native monoclinic insulin as a search model. A model built up from smaller units could help to avoid model bias, since the aggregation of the mutant could be slightly different. The molecular replacement solutions were very clear and produced a recognisable hexamer instantly. Indications of the positions for the new histidine side chains were present in the first electron density maps (see for example figure 3.22), and became clearer as refinement progressed.

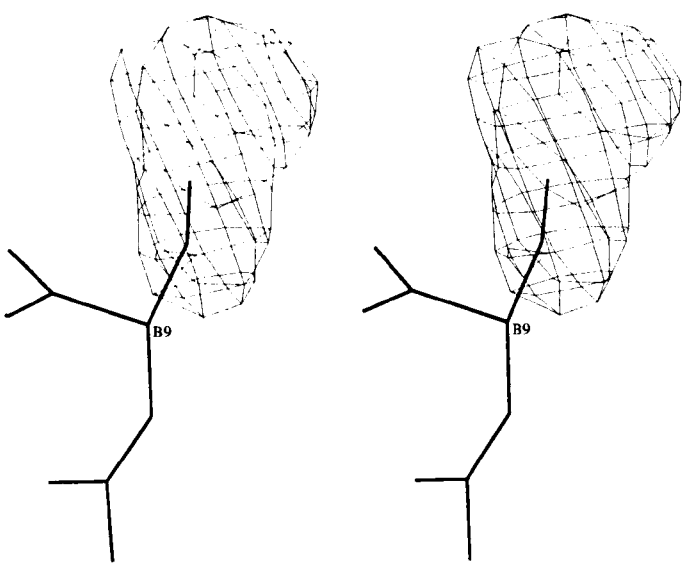

Figure 3.22: First indication of the mutation sites of B9H insulin, monoclinic form. Stereo view of one of the B9 residues with $F_{obs}-F_{calc}$ map at $+3\sigma$.

Omit maps calculated (as described in section 3.5.1) at the end of refinement confirm the final positions of the mutated residues (see figure 3.23). The six B9 histidines point towards the centre of the hexamer, as shown in figure 3.24.

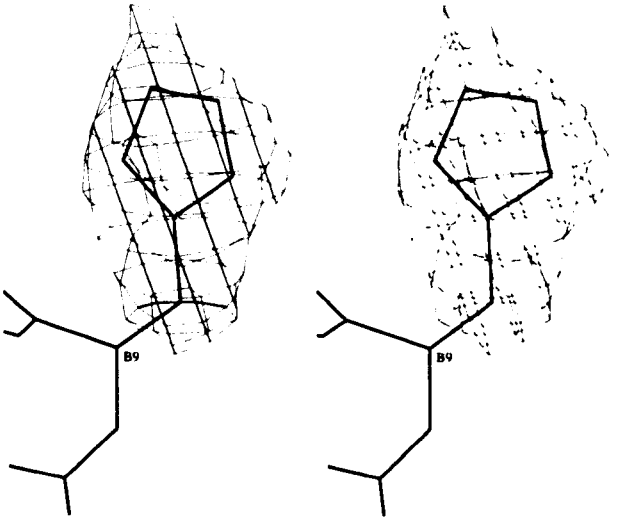

Figure 3.23: Omit maps for the mutation sites of B9H monoclinic insulin after refinement. Stereo view of the same B9 residue with $F_{obs}-F_{calc}$ map at $+3\sigma$ in the same orientation as figure 3.22.

During the course of refinement it became clear that apart from the two traditional zinc ions situated on the non-crystallographic threefold axis (ZN1 and ZN2), tetrahedrally coordinated by three B10 histidines and a chloride ion, more zinc was present in the structure. There is potential for zinc in the centre of the hexamer, coordinated by the Nϵ2 atoms of the B9 histidines, and by the other nitrogen atom (Nδ1) of the B9 histidines, also coordinated by Nδ1 of a B10 histidine of another monomer within the hexamer. Zinc was included in places where the electron density clearly indicated an atomic position, less than ~2.5Å away from a histidine nitrogen atom. This resulted in the addition of ten zinc ions, all with half occupancy. The zinc ions coordinated by two Nδ1 atoms from a B9 and a B10 histidine residue, were refined with two half occupied chloride ions as third and fourth ligands. The

Figure 3.24: Stereo view of the core of B9H monoclinic insulin

environments of all 12 zinc ions are tabulated in table 3.15. The zinc ions are included in figure 3.24, along with the B10 histidines and the B13 glutamic acids. The picture shows the conformation of the B13 residues, all pointing towards the centre of the hexamer. This would have been highly unlikely without the zinc cations to counterbalance the negative charges of the glutamic acids.

There is no indication of zinc at B9Nδ1 or F9Nδ1. As may be seen from table 3.15, not all zinc positions are equally well-defined. However, the environments are unsuitable for water molecules. The exact amount of zinc in the crystals can not be determined unambiguously by crystallographic means at this resolution.

The structural biology of zinc in relation to histidine has been reviewed by Christianson (1991). Although the zinc ions on the non-crystallographic threefold axis were refined with occupancy 1.0, full occupancy of any of the zinc ions is not warranted by the temperature factors, the electron density and the chemistry of zinc and histidines at the pH of the crystal (which is 7.27). The temperature factors of the 'axial' zinc ions is within the same region as those of the nitrogens coordinated to them (14.6-26.5Å^2), leading to the conclusion that the occupancy of those zinc ions must be very close to unity. The occupancies of the other ten zinc ions vary, but judged by their refined temperature factors and those of their ligands, these zinc ions have occupancies 0.5 or less.

zinc	tf[1](Å^2)	coordination	ligands (distance in Å)					
1	20.5	tetrahedral	D10Nϵ2	F10Nϵ2	L10Nϵ2	CL11		
			2.09	1.96	1.94	2.18		
2	19.5	tetrahedral	B10Nϵ2	H10Nϵ2	J10Nϵ2	CL21		
			1.96	2.08	2.05	2.16		
3	38.5	tetrahedral	B10Nδ1	H9Nδ1	CL31	CL32		
			2.13	2.12	2.26	2.14		
4	35.5	tetrahedral	F10Nδ1	D9Nδ1	CL41	CL42		
			2.12	2.09	2.12	2.63		
5	46.5	tetrahedral	D10Nδ1	L9Nδ1	CL51	CL52		
			2.09	2.07	2.65	2.06		
6	43.5	tetrahedral	H10Nδ1	J9Nδ1	CL61	CL62		
			2.09	2.09	2.43	1.87		
7	45.0	irregular	F9Nϵ2	F13Oϵ2	H13Oϵ2	HOH443	HOH605	
			2.05	2.20	1.67	2.74	2.23	
8	55.5	irregular	D9Nϵ2	D13Oϵ2	B13Oϵ2	ZN22	HOH443	HOH605
			2.12	2.68	2.52	2.89	2.73	2.59
9	55.5	irregular	H9Nϵ2	F13Oϵ1	F13Oϵ2	B13Oϵ1	B13Oϵ2	HOH725
			2.06	2.55	2.57	3.14	3.27	1.91
10	53.0	irregular	J9Nϵ2	H13Oϵ2	J13Oϵ2	L13Oϵ2	HOH443	HOH712
			2.08	3.00	2.80	2.25	2.67	2.43
11	56.0	irregular	L9Nϵ2	J13Oϵ2	L13Oϵ2	HOH443		
			2.09	2.70	2.42	2.66		
12	48.5	irregular	B9Nϵ2	B13Oϵ2	D13Oϵ2	ZN8	HOH443	HOH712
			2.53	2.63	2.12	2.89	2.70	2.53

Table 3.15: Environments of the zinc ions in monoclinic B9H insulin. Distances between zinc and histidine nitrogens were restrained to 2.1Å. Distances between zinc and chloride were not restrained. [1]tf = temperature factor.

In order for histidine to coordinate to two metal ions at the same time, it will have to lose both its protons. The pK_a of the second proton in histidine is 14.0-14.5, while this pK_a may be lowered around two pH units by coordination to a metal ion. In protein structures, this has been seen in copper-zinc superoxide dismutase (Tainer *et al.*, 1982). The pH of the present mutant crystallisation is much lower than the 12.0 seen as viable for binding of two metal ions to one histidine residue, so the local environment of these histidines has an exceptionally high effective pH. The locality of the zinc binding potential can also be derived from the geometry around the metal positions. The geometry of Zn^{2+} interactions with sp^2 nitrogen-containing heterocycles (as in histidine) has been examined by Vedani and Huhta (1990). The metal ion prefers a head-on and in-plane approach to the lone electron pair of the nitrogen atom. From their study of the small molecule structures containing zinc-imidazole centres in the Cambridge Structural Database (Allen *et al.*, 1979), Vedani

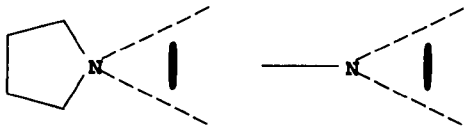

Figure 3.25: Preferred binding of zinc to histidine: the clustering of 'head-on' (left part of the figure) and 'in-plane' (right part of the figure) preference of zinc in histidine-environment in the Cambridge Structural Database indicates a maximum deviation of 30° in both cases. The figure shows a top and side view of a histidine ring, and a cone with a 30° opening angle in dashed lines. The zinc ions in the Cambridge Structural Database are all found in the area indicated by a black ellipse.

and Huhta have deduced that the zinc ion normally lies well within 30° from the C-N-C bisector (see figure 3.25). The geometry of some of the zinc ions in the monoclinic B9H mutant, however, deviates significantly from their findings (see table 3.16). The four zinc ions tetrahedrally coordinated by two histidines (B9 and B10, both by nitrogen Nδ1) and two chlorides are significantly out-of-plane. In an ideal zinc-histidine pair, the plane deviation would be zero and the angles would both be 126°. In the case of deviations from the histidine plane, 'head-on' only manifests itself in very similar CC and CN angles (for a definition of these angles, see the legend of table 3.16), decreasing with increased plane deviation. Therefore, the ZN1-L10Nϵ2 and ZN9-H9Nδ1 pairs are outliers.

zinc ion	nitrogen	plane deviation	angle CC †	angle CN ‡
ZN1	D10Nϵ2	6.7	120.0	125.9
	F10Nϵ2	4.7	127.4	123.6
	L10Nϵ2	21.0	148.4	103.2
ZN2	B10Nϵ2	10.0	139.0	118.5
	H10Nϵ2	3.4	120.3	108.6
	J10Nϵ2	13.5	139.2	110.8
ZN3	B10Nδ1	30.7	112.8	129.8
	H9Nδ1	74.0	103.0	97.0
ZN4	F10Nδ1	26.0	120.2	111.6
	D9Nδ1	62.3	123.2	103.6
ZN5	D10Nδ1	27.8	118.2	122.9
	L9Nδ1	66.7	112.8	93.4
ZN6	H10Nδ1	37.9	118.9	116.7
	J9Nδ1	71.8	107.9	93.0
ZN7	F9Nϵ2	18.4	129.1	113.1
ZN8	D9Nϵ2	16.8	114.8	138.2
ZN9	H9Nϵ2	22.4	145.5	103.8
ZN10	J9Nϵ2	11.4	136.3	114.8
ZN11	L9Nϵ2	13.6	109.1	138.5
ZN12	B9Nϵ2	10.8	122.8	126.9

Table 3.16: Geometry of the zinc ions in monoclinic B9II. Plane deviations and angles in °.
†angle CC = angle ZN-Nϵ2-Cδ2 or angle ZN-Nδ1-Cγ;
‡angle CN = angle ZN-Nϵ2-Cϵ1 or angle ZN-Nδ1-Cϵ1.

Phenol molecules could be identified in positions described by Derewenda *et al.* (1989) with hydrogen bonds to main chain oxygen and nitrogen in A6/A11, C6/C11, E6/E11, G6/G11, I6/I11 and K6/K11 all between 2.5 and 3.1Å. The refined positions of the phenol molecules are confirmed by omit maps (as described in section 3.5.1) at the end of refinement (see figure 3.26).

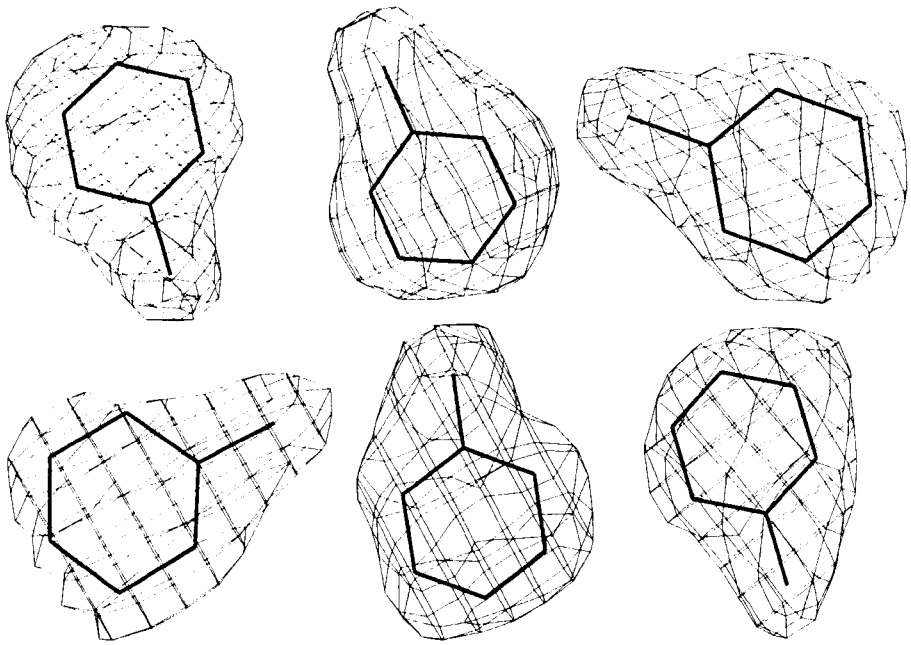

Figure 3.26: Omit maps for the phenol molecules in B9H monoclinic insulin after refinement. The spatial separation of the molecules in the figure does not reflect that in the model.

Although the first two residues of chains D, H and L are not clearly visible in electron density maps, it can be determined without doubt that the B9H mutant in monoclinic form is an example of an R6 hexamer. The temperature factors for the first two residues of chains D and H are relatively high (see figure 3.29), reflecting their ambiguity in electron density maps. The comparatively low temperature factor of residue L1 is due to the fact that it was reset but not refined (the residue was

excluded from further refinement). The N-termini of the B chains are all essentially helical, as can be seen in figure 3.27, with in some cases a small distortion at the end resulting in an extended conformation for the first residue.

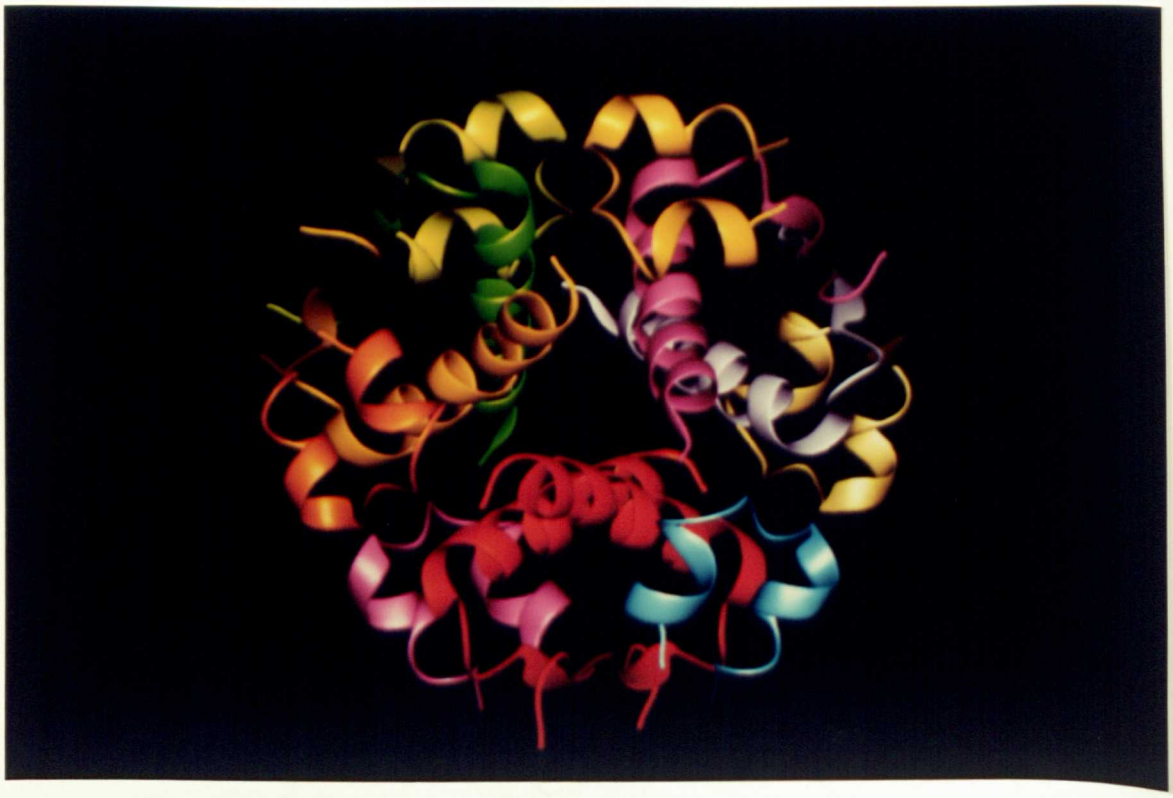

Figure 3.27: Ribbon diagram of B9H monoclinic insulin. All twelve chains in the hexamer in different colours.

The C-termini of the B chains are disordered in all six monomers, resulting in high temperature factors (see figure 3.29). Parts of these C-termini were excluded from refinement. The electron density for the A chains was easily interpretable and the temperature factors were fluctuating within the range also seen in other insulin structures (see figure 3.28).

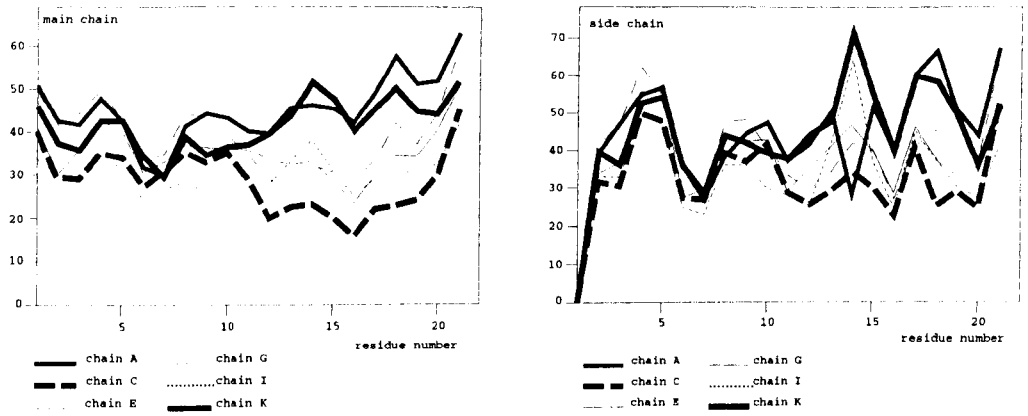

Figure 3.28: Average temperature factors per residue for the A chains of B9H insulin, monoclinic form

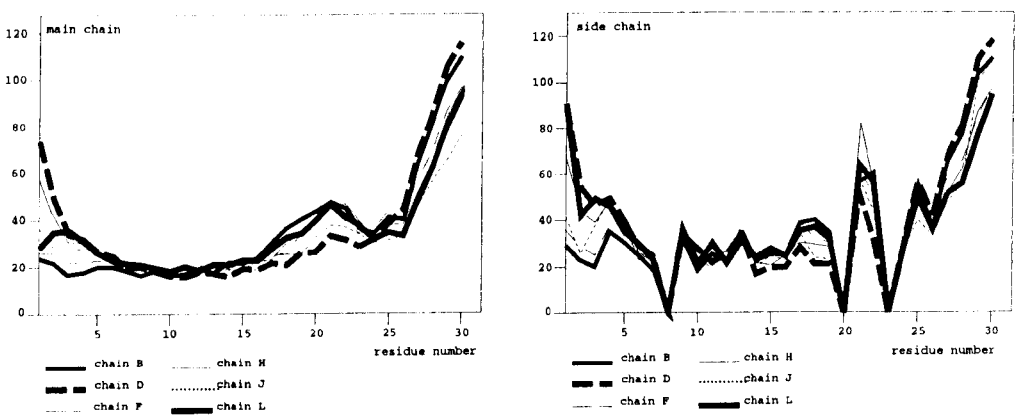

Figure 3.29: Average temperature factors per residue for the B chains of B9H insulin, monoclinic form

Rhombohedral crystal form

The B9H mutant crystallised readily in rhombohedral form (space group R3) in T3R3-conformation. Cryocrystallographic data collection was successful with a non-standard cryoprotectant (1,2-ethanediol), after which structure determination and refinement proceeded smoothly until problems were encountered with the restraint of special distances involving atoms at special positions and atoms from symmetry-related molecules. The mutant is isomorphous with native 4Zn insulin, and has zinc ions on the threefold axis, both in the T-conformation and in the R-conformation trimer (see figure 3.30). Both zinc ions are tetrahedrally coordinated by three B10 His Nϵ2 and a chloride anion.

During the course of refinement, the presence of two more zinc ions was firmly established. A third zinc ion is tetrahedrally coordinated by B5 His Nϵ2 and B9 His Nϵ2 (both from molecule 2) and two water molecules (see figure 3.31). The temperature factor of this zinc ion is slightly higher than those of its ligand atoms: 45Å^2 for the zinc ion and 31 to 39Å^2 for the ligand atoms. A fourth zinc ion is tetrahedrally coordinated by B10 His Nδ1 from molecule 1, B9 His Nδ1 from a symmetry-related molecule 1, and two water molecules (see figure 3.32). The geometry of the B9 histidine in relation to the zinc ion, is imperfect, but can not be restrained in this particular refinement program because it is a symmetry-related residue. This histidine residue may be disordered, which would explain the presence of B13 Glu from molecule 2 within hydrogen bonding distance if the histidine ring is in a slightly different orientation. The occupancy of the zinc ion was set at 2/3, after which the temperature factor refined to 34Å^2. The ligands for this zinc ion have temperature factors between 28 and 42Å^2. If the occupancy of this zinc ion is assumed to be correctly assigned a value of 2/3, the stoichiometry of zinc to B10 histidine (molecule 1) is 1:1. This means that all histidines in the structure lose only

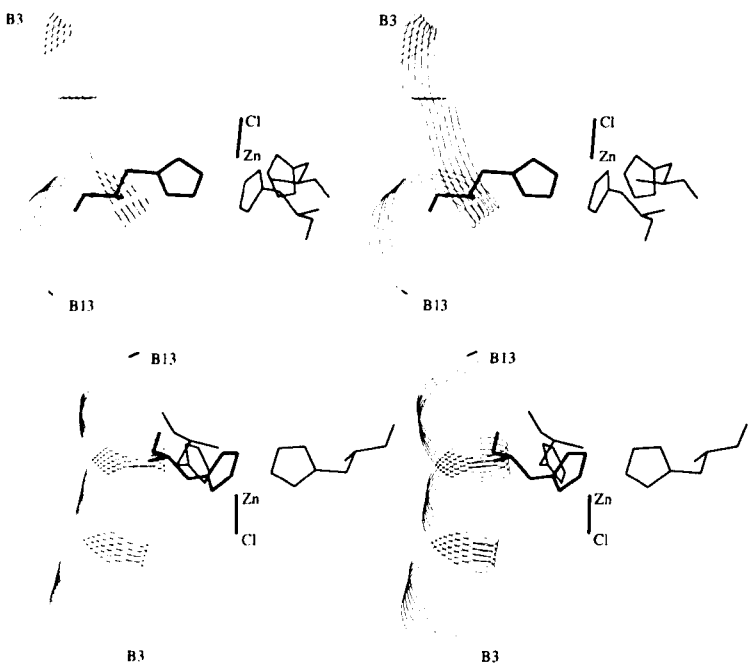

Figure 3.30: Axial zinc ions in B9H rhombohedral insulin. Stereo views of both axial zinc ions; top: situation in molecule 1, T-conformation; bottom: situation in molecule 2, R-conformation. Residues B10 from both molecules shown in thick lines, with symmetry-related B10 residues in thinner lines. The spatial separation of the zinc ions in the figure does not reflect that in the model.

one proton, which is the more appropriate at the pH of the crystallisation (which is 7.43).

The geometry of the zinc ions in rhombohedral B9H insulin with respect to the histidine rings is reasonable compared to the averages found by Vedani and Huhta (1990), for the axial zinc ions and the zinc ion coordinated by histidines B5 and B9 from molecule 2 (see table 3.17). The fourth zinc ion, however, is distinctly out-of-plane, because the position of the B10 histidine of molecule 1 is fixed by the axial zinc ion on the other nitrogen. The geometry of the tetrahedral coordination of the fourth zinc ion is, therefore, distorted, while that of the other extra zinc ion (ZN3) is regular, as shown in table 3.18.

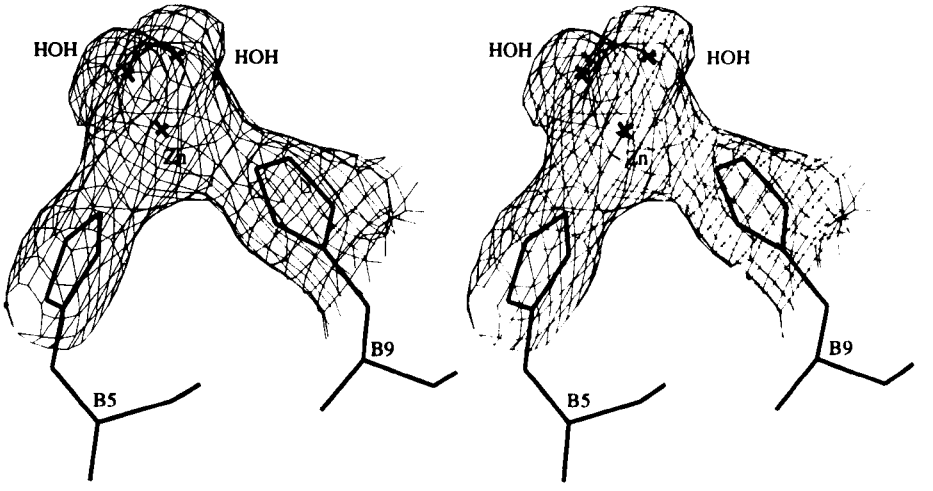

Figure 3.31: Extra zinc in molecule 2 in B9H rhombohedral insulin. Stereo view of the tetrahedral coordination by B5, B9 and two water molecules, with $2F_{obs}-F_{calc}$ electron density at $+1\sigma$.

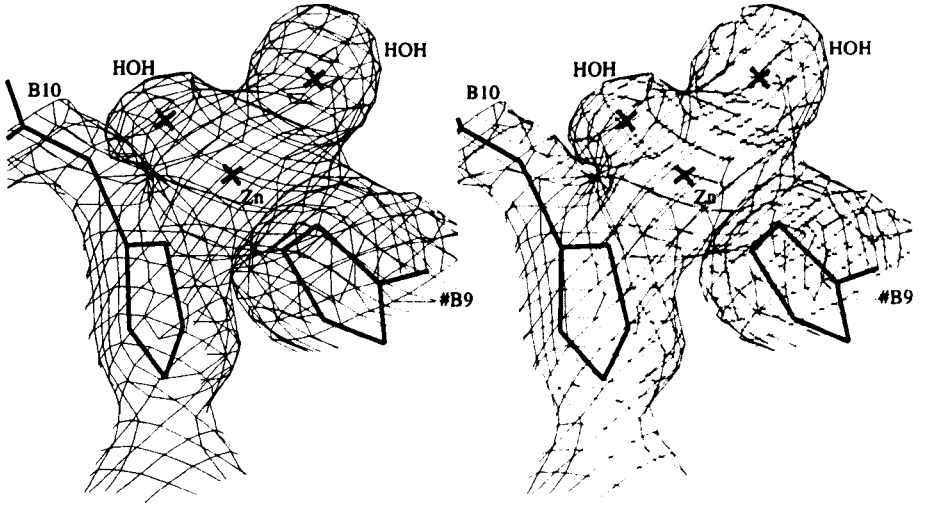

Figure 3.32: Extra zinc in molecule 1 in B9H rhombohedral insulin. Stereo view of the tetrahedral coordination by B10, a symmetry-related B9 and two water molecules, with $2F_{obs}-F_{calc}$ electron density at $+1\sigma$.

zinc ion	nitrogen	plane deviation	angle CC †	angle CN ‡
ZN1	B10Nϵ2	16.5	110.2	139.1
ZN2	D10Nϵ2	11.7	114.5	136.4
ZN3	B9Nϵ2	16.1	114.3	131.6
	B5Nϵ2	14.1	133.1	118.7
ZN4	D10Nδ1	27.0	145.0	103.6
	#D9Nδ1	34.9	121.8	112.3

Table 3.17: Geometry of the zinc ions in rhombohedral B9H. Plane deviations and angles in °.
†angle CC = angle ZN-Nϵ2-Cδ2 or angle ZN-Nδ1-Cγ;
‡angle CN = angle ZN-Nϵ2-Cϵ1 or angle ZN-Nδ1-Cϵ1.

ligand 1	zinc ion	ligand 2	tetrahedral angle (°)
W10	ZN3	W11	107.8
W10	ZN3	B9Nϵ2	103.0
W10	ZN3	B5Nϵ2	113.4
B9Nϵ2	ZN3	B5Nϵ2	107.1
B5Nϵ2	ZN3	W11	114.2
B9Nϵ2	ZN3	W11	110.8
W3	ZN4	W4	110.8
W3	ZN4	D10Nδ1	113.6
W3	ZN4	#D9Nδ1	97.3
D10Nδ1	ZN4	#D9Nδ1	123.2
#D9Nδ1	ZN4	W4	101.1
D10Nδ1	ZN4	W4	109.7

Table 3.18: Tetrahedral environment for the extra zinc in rhombohedral B9H insulin.

The conformation of the N-terminus of the B chain of molecule 2, although considered to be in the R-state, is not entirely helical. Like in the B13Q mutant, the side chain of residue B3 occupies the position of the main chain of residues B1 and B2 if the helix would have extended to B1, and *vice versa*. This part of the structure is shown in figure 4.2. The temperature factors for the B9H rhombohedral insulin mutant are shown in tables 3.19 and 3.20. The temperature factors for the residues in the A chains do not deviate much from the average for the whole chain. Both B chains show similar patterns: low temperature factors for the helical residues and residues 24-26 of the β-sheet, higher temperature factors for the loop at 20-23 and towards the N-terminus, and high temperature factors for the C-termini indicating disorder.

	chain A		chain C	
residue	main chain(Å^2)	side chain(Å^2)	main chain(Å^2)	side chain(Å^2)
1	56.8	N/A	40.5	N/A
2	52.3	47.1	41.0	37.0
3	56.2	54.9	42.6	43.7
4	58.3	63.3	47.0	55.2
5	55.5	58.8	47.2	52.9
6	50.0	46.1	41.4	45.6
7	54.3	49.3	45.5	44.4
8	63.4	66.4	54.3	57.2
9	61.5	63.9	57.2	58.2
10	53.2	54.0	57.1	58.3
11	45.6	41.7	57.7	56.8
12	37.0	38.7	57.1	57.1
13	33.9	41.4	53.6	53.7
14	35.2	36.7	50.5	50.0
15	35.2	37.3	48.3	58.3
16	35.4	39.1	41.4	43.7
17	35.9	45.9	35.3	47.6
18	39.5	39.6	35.2	36.1
19	34.8	38.5	30.9	33.0
20	35.8	33.6	35.0	33.5
21	48.6	57.1	42.9	40.8
total average	46.592	46.948	45.800	47.422

Table 3.19: Average temperature factor per residue of A chain of B9H insulin, rhombohedral form

residue	chain B		chain D	
	main chain(Å^2)	side chain(Å^2)	main chain(Å^2)	side chain(Å^2)
1	50.0	50.0	46.5	43.7
2	56.4	50.0	46.1	47.1
3	58.8	63.5	52.2	58.2
4	47.3	56.3	48.6	54.2
5	35.7	38.0	44.7	52.5
6	30.4	35.2	40.1	40.9
7	33.4	41.4	33.9	40.3
8	28.6	N/A	31.2	N/A
9	29.6	37.0	25.7	40.8
10	25.7	25.7	25.9	29.0
11	27.4	33.8	27.3	32.5
12	24.9	28.0	26.1	27.4
13	25.1	41.6	21.6	44.4
14	26.8	26.7	24.1	20.9
15	26.1	29.4	27.0	26.0
16	25.3	27.7	27.8	28.3
17	26.8	30.0	29.5	34.4
18	32.7	34.3	34.7	34.7
19	34.6	26.9	40.7	31.5
20	39.5	N/A	48.1	N/A
21	43.3	38.0	56.0	62.5
22	44.1	52.7	54.6	58.4
23	38.0	N/A	41.3	N/A
24	32.7	29.1	30.0	29.3
25	35.1	39.4	30.3	33.8
26	39.5	37.0	28.1	34.0
27	55.0	55.7	39.1	36.2
28	66.9	64.3	49.6	44.8
29	77.9	78.0	65.0	66.6
30	83.8	85.1	72.7	74.8
total average	40.042	42.527	38.947	42.301

Table 3.20: Average temperature factor per residue of B chain of B9H insulin, rhombohedral form

3.6.3 B8 Gly→Ser, B13 Glu→Gln, B30 Thr-NH_2 human insulin

At the neutral pH of the crystallisation, the B8S/B13Q/B30amide mutant forms stable hexamers, in R6-state because of the addition of phenol. Although the presence of a hydrophilic serine residue in place of glycine at position B8 causes unexpectedly increased hydrophobicity (characterised by a high capacity factor relative to pig insulin in elution from the HPLC column (Markussen *et al.*, 1987)), the conformation of residues B1-B19 is entirely helical without much distortion (see figure 3.33). All six B8 mutation sites are clear in electron density, and the serines make hydrogen

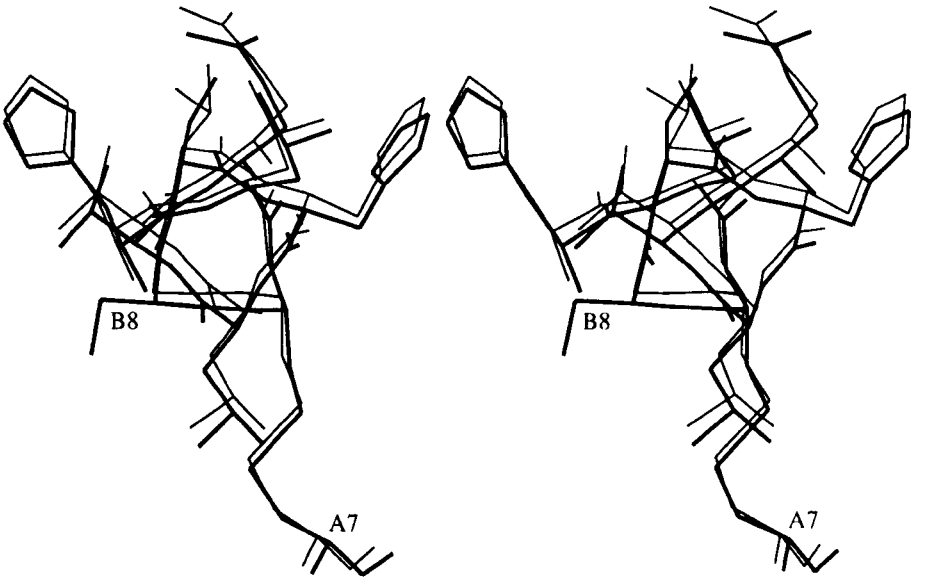

Figure 3.33: Stereo view of the overlap of the α-helix at B1-B19 in B8S/B13Q/B30amide and native monoclinic insulin.
Mutant structure in thick lines, native structure in thin lines.

bonds mainly with B26 tyrosine Oη and B4 main chain oxygen of the same chain. The increased hydrophobicity of the mutant can not be explained from the present crystal structure. The presence of an intricate hydrogen bonding network of the B13 glutamine residues in the mutant was investigated. However, five out of six of the glutamine side chains bend away from the centre of the molecule, and hydrogen bond to B9O and B9Oγ of the other monomer in the same dimer. Only F13 points straight towards a central water molecule, hydrogen bonding to it. L13 bends away from the centre, but not quite towards J9, hydrogen bonding to two water molecules instead. There are no hydrogen bonds between the glutamine residues. Thus, the 'ring' structure of six glutamines as proposed by Markussen *et al.* (1987), is not observed.

Two zinc ions are clearly identifiable on the non-crystallographic threefold axis of the hexamer. ZN1 is tetrahedrally coordinated by Nϵ2 of histidines D10,

F10 and J10 in the present structure, with chloride as a fourth ligand. ZN2 is tetrahedrally coordinated by Nϵ2 of histidines B10, H10 and L10, and a chloride ion. Phenol molecules are observed in positions as described by Derewenda *et al.* (1989), hydrogen bonded to A6O and A11N of every A chain in the hexamer. The refined positions of the phenol molecules are confirmed through omit maps (calculated as described in section 3.5.1) at the end of refinement (see figures 3.34 a to f).

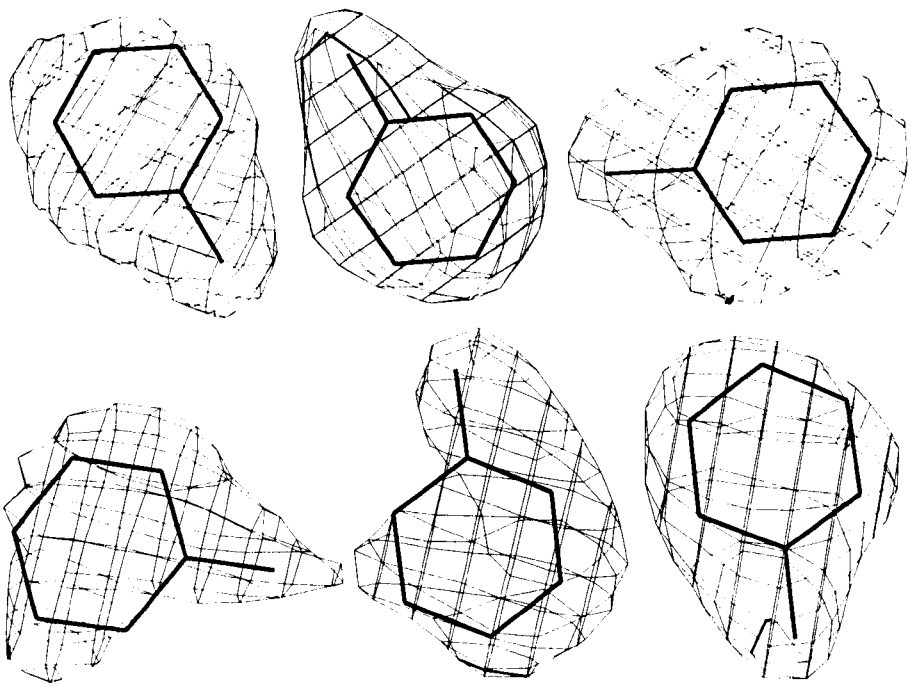

Figure 3.34: Omit maps for the phenol molecules of B9S/B13Q/B30amide insulin after refinement. The spatial separation of the molecules in the figure does not reflect that in the model.

The chain termini of the B8S/B13Q/B30amide mutant are as follows:

- Both A chain termini are clear in electron density maps;
- The B chain N-terminus is helical to B1 and clearly identifiable in electron density maps. The C-terminus is disordered beyond B28;

- Both C chain termini are clear;
- The D chain N-terminus is poorly defined but the geometry of the final model is good. The C-terminus is disordered beyond D28;
- The E chain N-terminus is unclear and its geometry is unfavourable. Repeated omit maps do not indicate better positions. The C-terminus is very clear. There is a salt bridge between one of the C-terminal oxygens of E21 and F22 Arg Nη_2;
- The F chain N-terminus is helical to F1 and clear in electron density. The C-terminal residues beyond F28 were not included in the refinement since there is no indication of their positions;
- The G chain N-terminus is well-defined. The last peptide bond in the C-terminal region has poor electron density;
- The H chain N-terminus is unclear, resulting in residue H1 being excluded from refinement. The C-terminus is well-defined apart from the side chain of H29; this residue has therefore been refined with zero occupancy for its side chain;
- Both I chain termini are clear in electron density;
- The J chain N-terminus is not completely clear in the electron density maps, but the geometry of the refined positions is good. The C-terminus is interpretable to the end, apart from the lysine side chain of J29, which is excluded from refinement;
- Both K chain termini are clear in the electron density maps;
- The L chain N-terminus has poor electron density, and seems to come into contact with the equally poorly defined N-terminus of the H chain. The C-terminal residues beyond L28 were excluded from refinement.

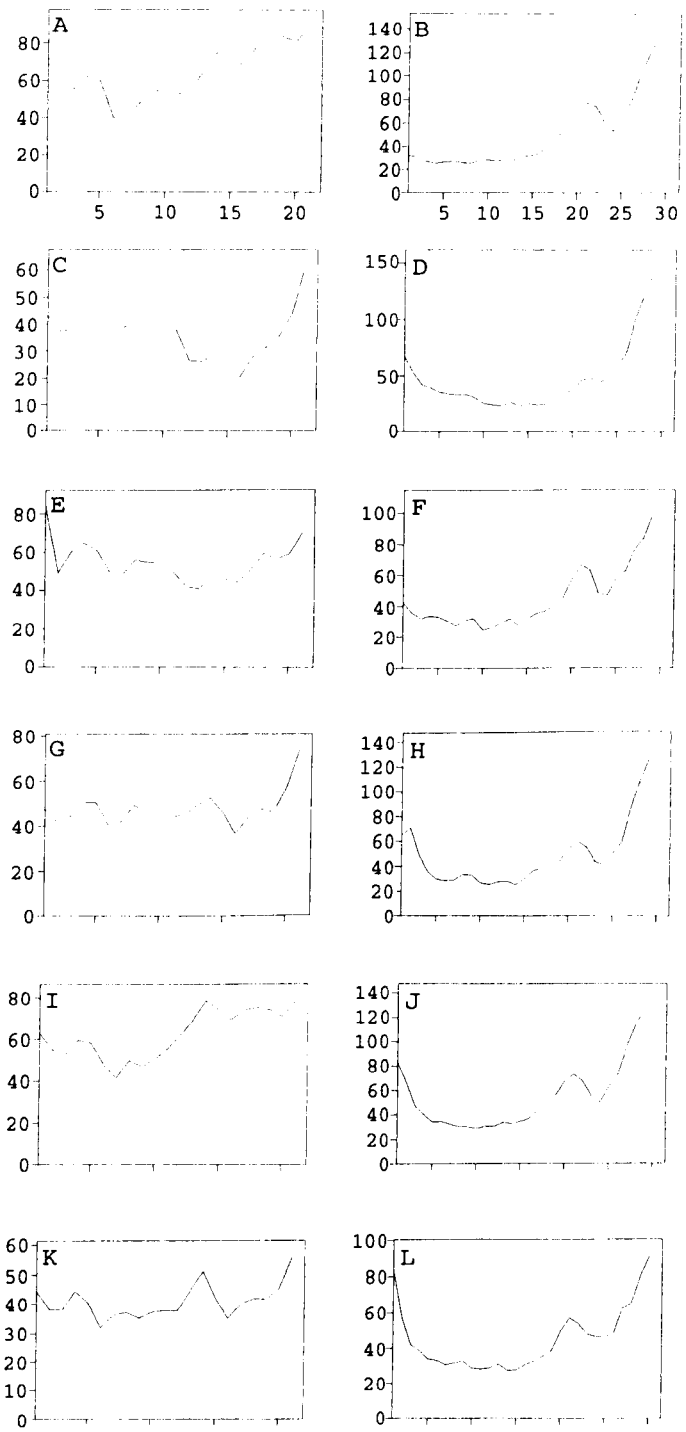

Figure 3.35: Average temperature factors per residue for the main chains of B8S/B13Q/B30amide insulin

Because the C-termini of the B chains are especially flexible, from the crystal structure of the B8S/B13Q/B30amide insulin no information is gained about the amide group of residue B30. More indications of the quality of the chain termini can be derived from the temperature factor plots, figures 3.35 a to l.

3.6.4 $Zn^{2+} \rightarrow Co^{2+}$ human insulin

The concentration of Co^{2+} ions in blood plasma is ten times lower than that of Zn^{2+} ions (0.002 mM and 0.02 mM, respectively (Glusker, 1991)). Although cobalt is not naturally found in insulin, the ions are similar in size and are expected to be able to take up the same role in insulin hexamerisation.

The crystallisation solution contained two reagents which have a drastic and almost antagonistic effect on the process of potential cobalt binding to insulin. On the one hand phenol promotes R-conformation. This might force Co^{2+} into a tetrahedral environment, which is unfavourable for d^7 ions like Co^{2+}. On the other hand halide ions are added, which allows Co^{2+} in tetrahedral coordination, as seen in some small molecule structures in the Cambridge Structural Database (Allen *et al.*, 1979). Co^{2+} insulin turned out to crystallise in rhombohedral space group R3, in T3R3-conformation. The disorder of B10 histidine (molecule 2) is clearly absent in Co^{2+} insulin, resulting in only two well-defined metal positions on the rhombohedral threefold axis. In molecule 2, the R-conformation trimer, Co^{2+} is tetrahedrally coordinated by three symmetry-related B10 His Nϵ2 and an iodide anion (see figure 3.36). In molecule 1 Co^{2+} is octahedrally coordinated by the other three symmetry-related B10 His Nϵ2 in the hexamer, and three symmetry-related water molecules (see figure 3.37).

A phenol molecule is present in the phenol pocket in the R3 trimer of Co^{2+} insulin, similar to that seen in monoclinic insulin, hydrogen bonded to A6O and A11N in molecule 2, confirmed by omit maps (as described in section 3.5.1) at the

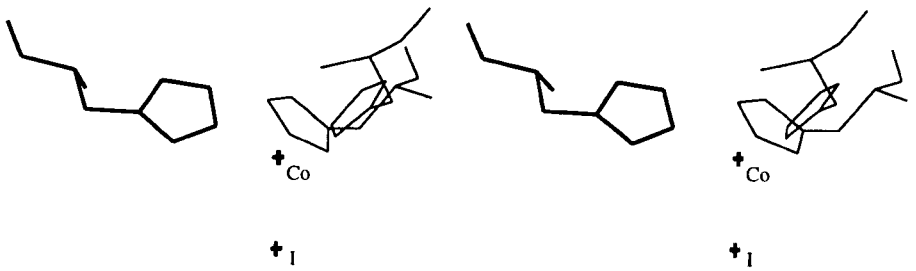

Figure 3.36: Cobalt ion in molecule 2 in CoI insulin. Stereo view of the tetrahedral coordination by B10 His (thick lines), two symmetry-related B10 His residues (thinner lines) and an iodide ion.

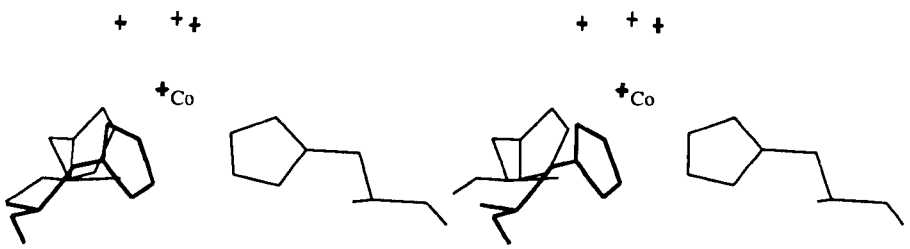

Figure 3.37: Cobalt ion in molecule 1 in CoI insulin. Stereo view of the octahedral coordination by B10 His (thick lines), two symmetry-related B10 His residues (thinner lines) and three symmetry-related water molecules.

end of refinement. This is the region of the off-axial zinc site in the R3 trimer in 4Zn native rhombohedral insulin, with which Co^{2+} insulin is isomorphous. A second phenol molecule is found in the same region of the molecule, with hydrogen bonds to B10 Nδ1 and B13 Oϵ1. The phenol molecules prevent disorder of B10 His, stabilising it in coordination with cobalt on the threefold axis (see figure 3.38).

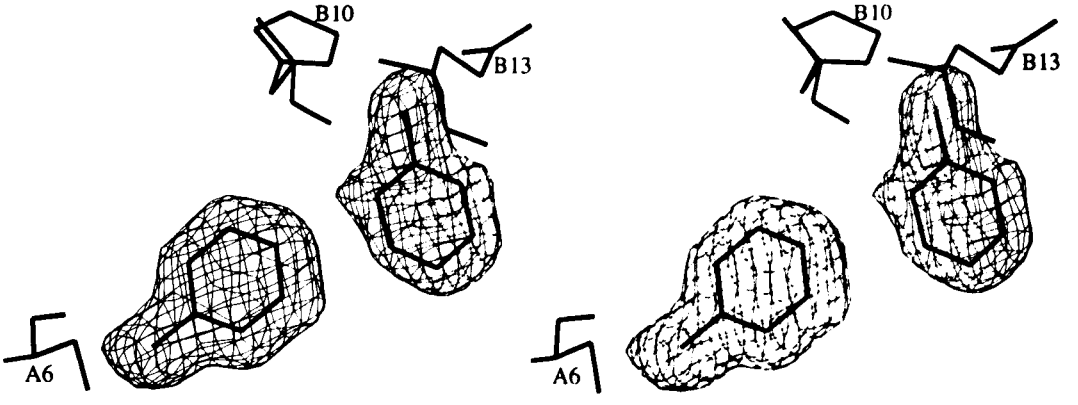

Figure 3.38: Phenol helps stabilise cobalt coordination in molecule 2 of Co^{2+} insulin. Stereo view of the two phenol molecules, hydrogen bonding to residues A6, B10 and B13; $F_{obs}-F_{calc}$ electron density at $+3\sigma$. The phenol molecules were omitted from these map calculations.

The N-terminus of the B chain in molecule 2, supposedly in R-conformation, is not entirely helical to residue B1. As seen with the B13Q and the B9H mutants, the first two residues of the B chain exhibit an extended conformation, in which not the main chain of residues B1 and B2 forms the start of the helix, but the side chain of residue B3. Although the temperature factors for the first residues of the B chain are rather high (see figure 3.39), the Ramachandran angles are much better in this conformation, and the electron density is convincing (see figure 4.2). The C-termini of both B chains have high temperature factors, reflecting their flexibility.

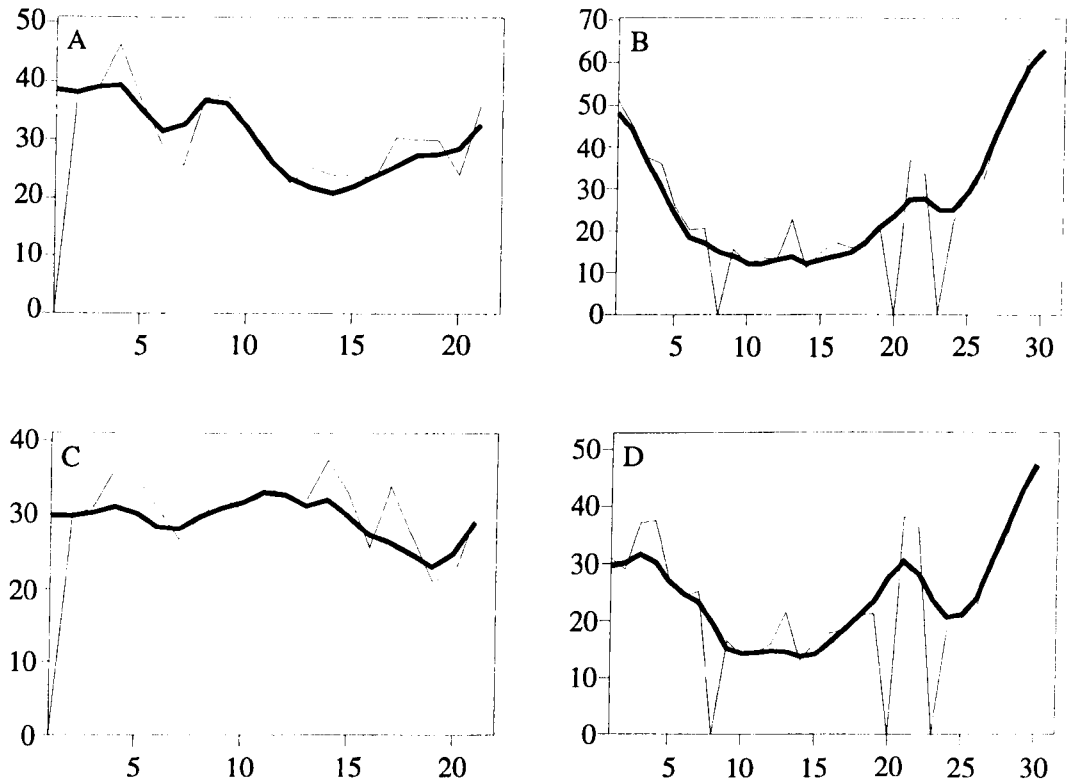

Figure 3.39: Average temperature factors per residue for CoI insulin. Main chains in thick lines, side chains in thin lines.

3.6.5 B9 Ser→Asp, B27 Thr→Glu human insulin

The B9D/B27E mutant is monomeric in solution (Brange *et al.*, 1988). The mutations in B9D/B27E insulin prevent its aggregation into hexamers because of the charge repulsion in the region where dimers would pack together into a hexamer. Without any added ions to balance the charges, the mutant does, however, crystallise as dimers isomorphously with orthorhombic native insulin. Although the pH of the crystallisation is unclear, it should be close to 6 since that is the pH of the buffer used. If the carboxylic acids at B9 and B13 have hydrogen bonds with each other, the pH must have been lower in order to allow partial protonation.

The non-crystallographic axis of the dimer is in the xz-plane, making a 30° angle with the x-axis. The packing of the mutant is shown in figure 3.40, with the direction of the dimer axis indicated.

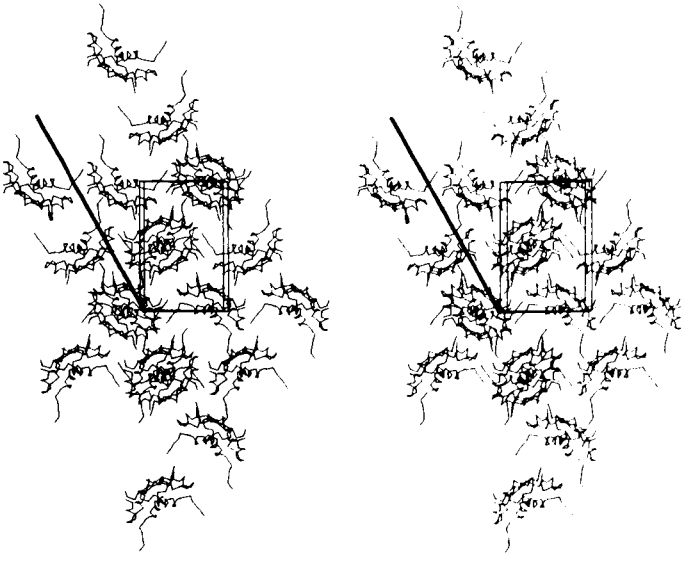

Figure 3.40: Packing of B9D/B27E insulin. Stereo view of the xz-plane, with the direction of the dimer axis indicated by a thick line. Cell edges also indicated.

Because the data for the B9D/B27E mutant were not optimal (see section 3.5.5), the maps were not convincing at every stage of the refinement. In the first maps, with the side chains at the mutation sites omitted (calculated as described in section 3.5.1), the electron density for the mutations at positions B9 was reasonably clear (see figure 3.41). For residue B27 of molecule 2, however, there was no indication as to where the side chain might be, while for the same site in molecule 1 there were problems with the positioning of the side chain with respect to residues B25 from both molecules (see figure 3.42). After a few refinement cycles with REFMAC the situation at the B27 site in molecule 1 became much clearer. Omit maps for the final positions of the mutated residues B27 are shown in figure 3.43. From figure 3.44 it can be seen that the B25 Phe side chains are aligned, and the B27 side chain of molecule 2 is involved in hydrogen bonding with four water molecules and a long contact (3.4Å) with the N-terminal nitrogen of the A chain of molecule 1 of a symmetry-related dimer. The B27 side chain of molecule 1 is apparently not solvated, and has one hydrogen bond with the B29 main chain oxygen of the same chain.

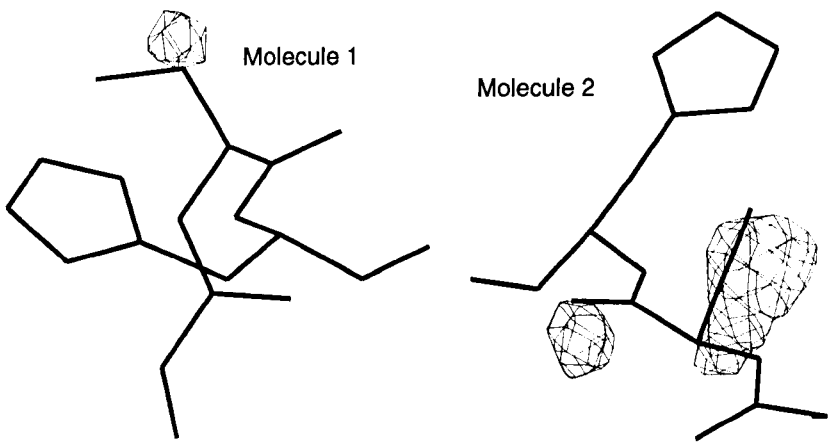

Figure 3.41: Omit map of B9 in both molecules of B9D/B27E insulin before refinement. $F_{obs}-F_{calc}$ electron density at $+3\sigma$.

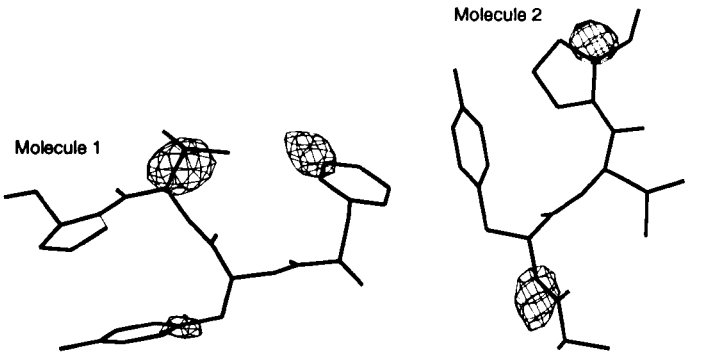

Figure 3.42: Omit map of B27 in both molecules of B9D/B27E insulin before refinement. $F_{obs}-F_{calc}$ electron density at $+3\sigma$.

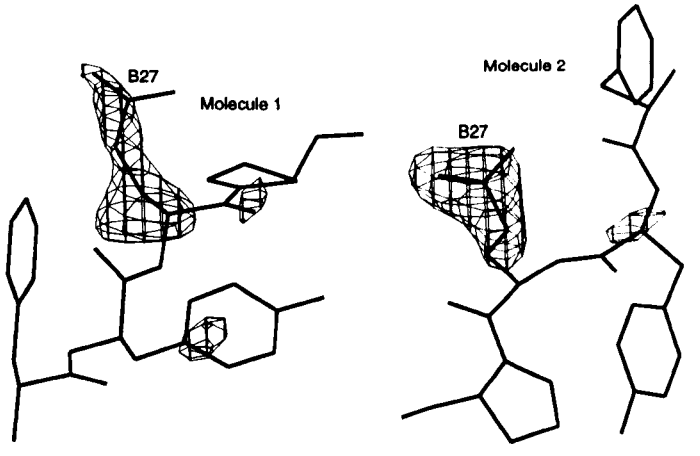

Figure 3.43: Omit map of B27 in both molecules of B9D/B27E insulin after refinement. $F_{obs}-F_{calc}$ electron density at $+3\sigma$.

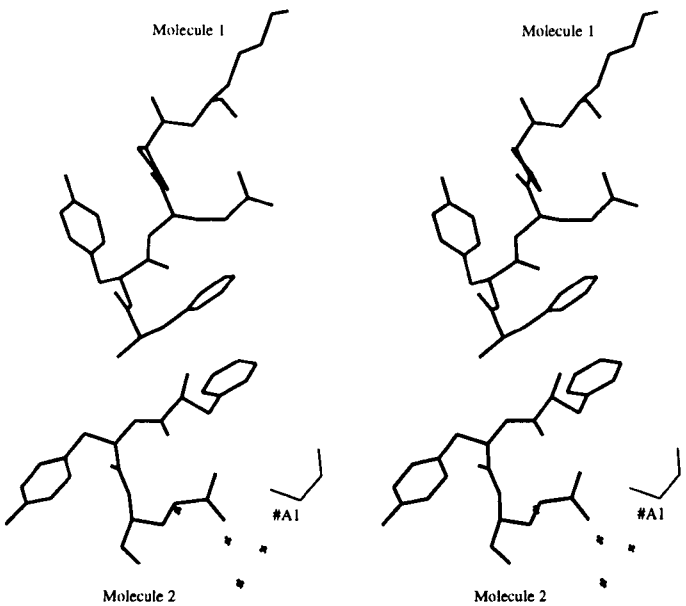

Figure 3.44: The situation around B25 and B27 in both molecules of B9D/B27E insulin. Stereo view including all atoms within 3.5Å of B25 and B27 residues.

The situation around the four carboxylic acid residues B9 and B13 is unclear. In an omit map of the final coordinates, the difference density for the aspartate and glutamate side chains is limited (see figure 3.45). Two conformations were modelled for the side chain of B13 in molecule 2, both with half occupancy. The five carboxylic groups as they were modelled all have hydrogen bonds, to each other and to water molecules. This indicates that the acid groups must have been at least partially protonated.

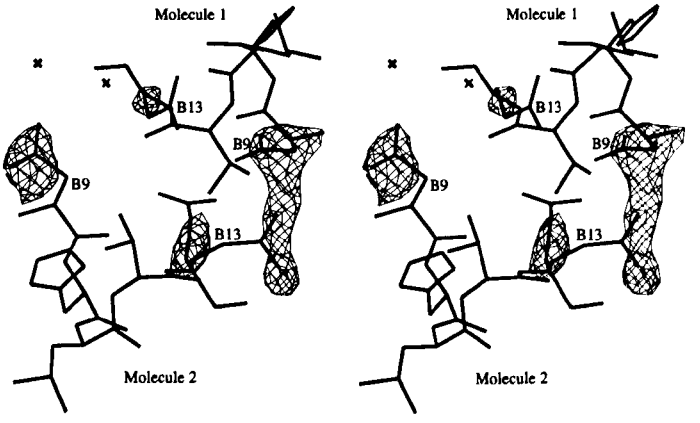

Figure 3.45: Omit map for residues B9 and B13 in both molecules of B9D/B27E insulin after refinement. Stereo view including two nearby water molecules; $F_{obs}-F_{calc}$ electron density at $+3\sigma$.

The temperature factors for the main and side chains of B9D/B27E insulin are shown in figure 3.46. Those of the first three residues of the D chain (molecule 1) have been reset to 20Å^2 and then not refined. The C-termini of the B chains are very flexible and thus have high temperature factors, but could be identified in electron density maps, apart from the side chains of both B29 lysines. The helical regions of the B chains have lower values, while the loops at B20 to B23 have slightly raised temperature factors, as expected.

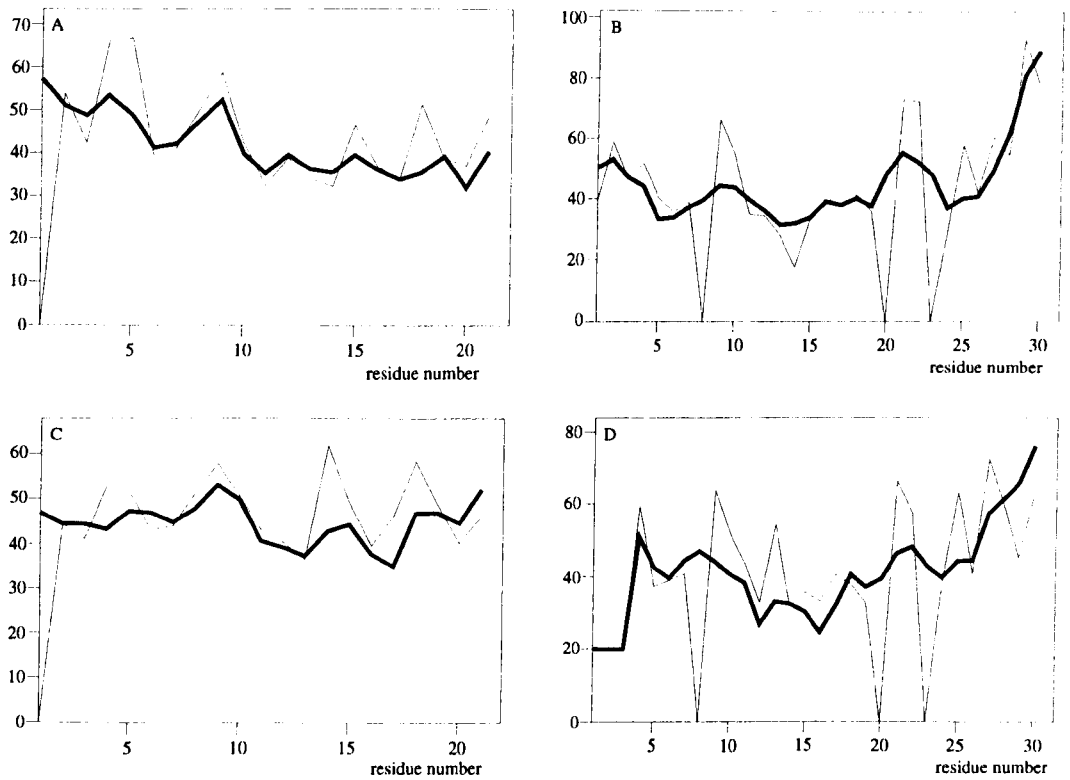

Figure 3.46: Average temperature factors per residues for B9D/B27E insulin. Main chains in thick lines, side chains in thin lines

3.6.6 A21 Asn→Gly, B9 Ser→Glu, B10 His→Glu human insulin

Aggregation of A21G/B9E/B10E insulin is hampered by charges at the surface where dimers would make contact to form a hexamer, and by removal of the zinc binding site. The mutant does, however form dimers in the usual way. The mutant crystallised in space group $P2_12_12_1$, with a packing different from that of native orthorhombic insulin. The non-crystallographic twofold axis of the mutant dimer is parallel to the x-axis. The packing of the mutant, indicating the screw axes of the space group and the direction of the non-crystallographic axis of the dimer, is shown in figure 3.47.

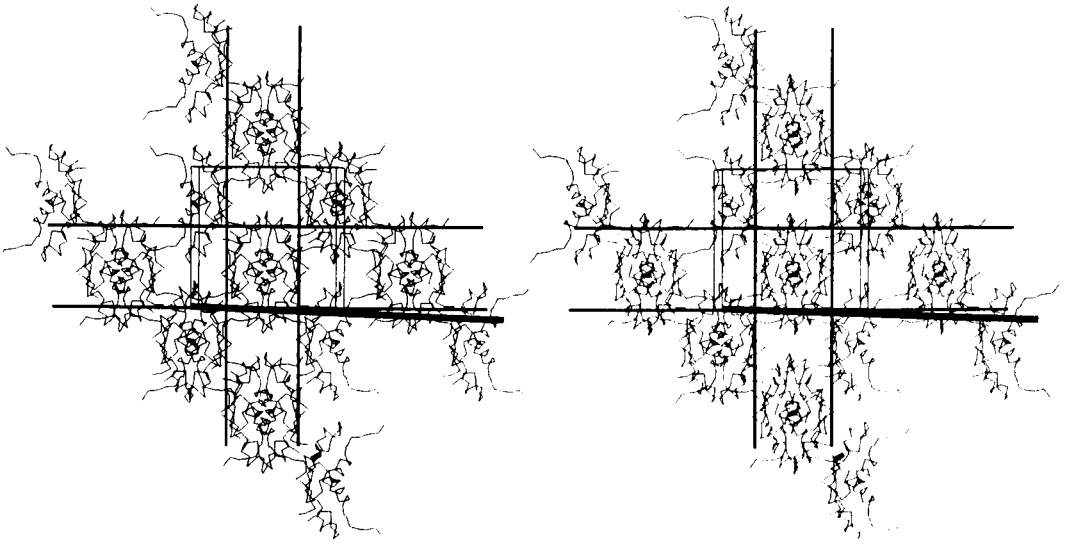

Figure 3.47: Packing of A21G/B9E/B10E insulin. Stereo view of the xz-plane, with the 2_1 screw axes indicated in thin lines, and the direction of the dimer axis as a thick line, almost parallel to one set of screw axes. Cell edges also indicated.

The most significant changes in the mutant with respect to the native dimer involve the packing of the C-terminus of the B chain of molecule 1 and the B20-B23 loop of molecule 2 with respect to each other, and with respect to the C-terminus of the B chain of molecule 2 in a symmetry-related dimer. In the starting model, these regions were in too close contact, but the first maps indicated improved positions, especially for the main chain of B27-B30 of molecule 1 (see figure 3.48). The refined positions of B27-B30 of molecule 1, B20-B23 of molecule 2 and B27-B30 of the appropriate symmetry-related molecule 2 are shown in figure 3.49, together with the starting positions for comparison.

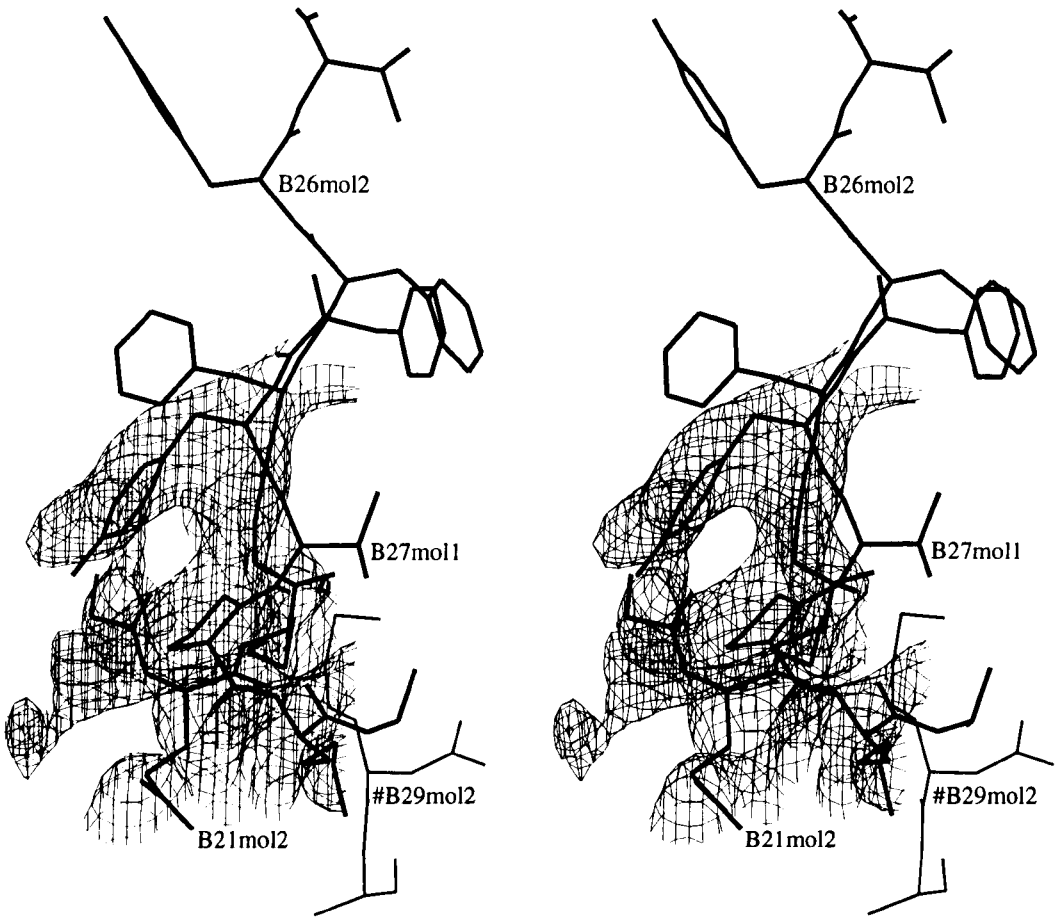

Figure 3.48: Indications for improved positions for the C-termini of the B chains of A21G/B9E/B10E insulin. Stereo view of the $2F_{obs}-F_{calc}$ map at $+1\sigma$ indicating a better position for residues B27 and B28 of molecule 1, to avoid clash with B29 from a symmetry-related molecule 2. Symmetry-related atoms in thinner lines.

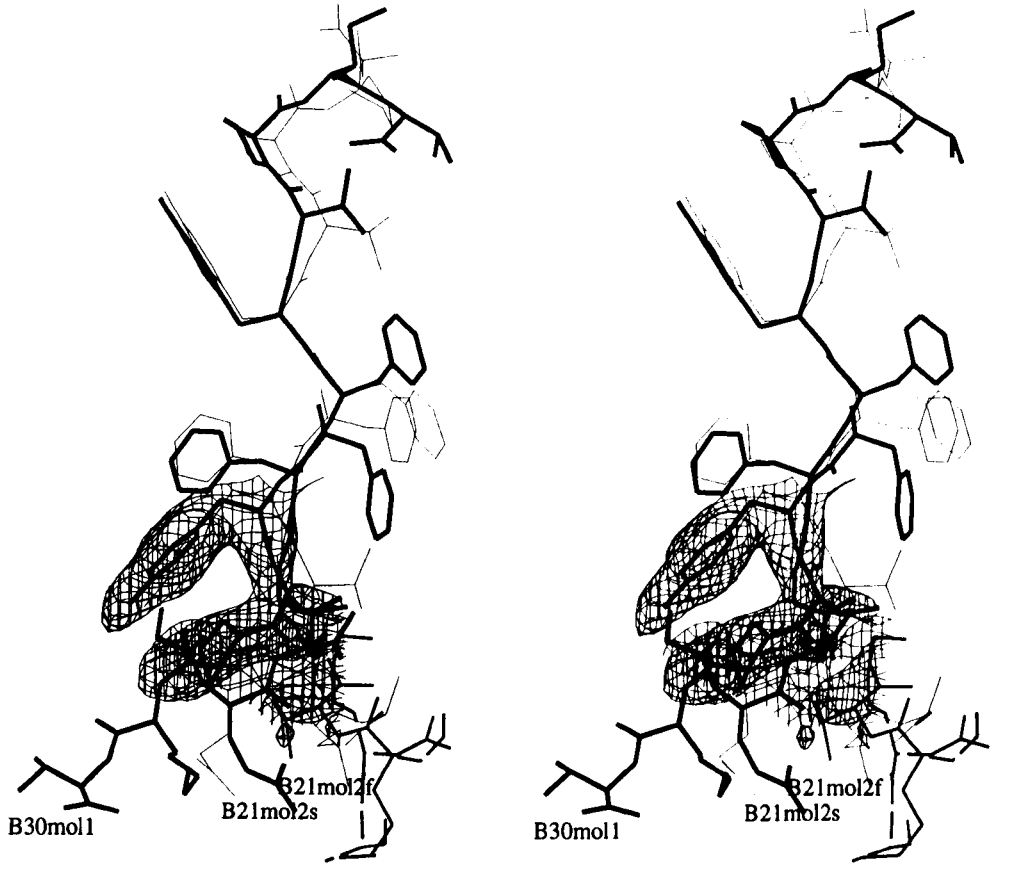

Figure 3.49: Final positions for the C-termini of the B chains of A21G/B9E/B10E insulin. Stereo view in the same orientation as figure 3.48, with $2F_{obs}-F_{calc}$ map at $+1\sigma$. Starting model (indicated by suffix 's' in molecule indicator) in thin lines, final model (indicated by suffix 'f' in molecule indicator) in thick lines. Symmetry-related atoms from both starting (dashed lines) and final (thin lines) models in bottom right-hand corner.

On the dimer surface where hexamerisation usually takes place, the mutant has six glutamic acid groups. At the pH of the crystallisation, which is 6.57, these are probably fully ionised and can not hydrogen bond with each other. Since the resolution of the data is only 2.8Å, the final omit maps (calculated as described in section 3.5.1) are somewhat poor (see figure 3.50). However, the indications for the positions of the refined side chains are present. Only two of the six glutamates make

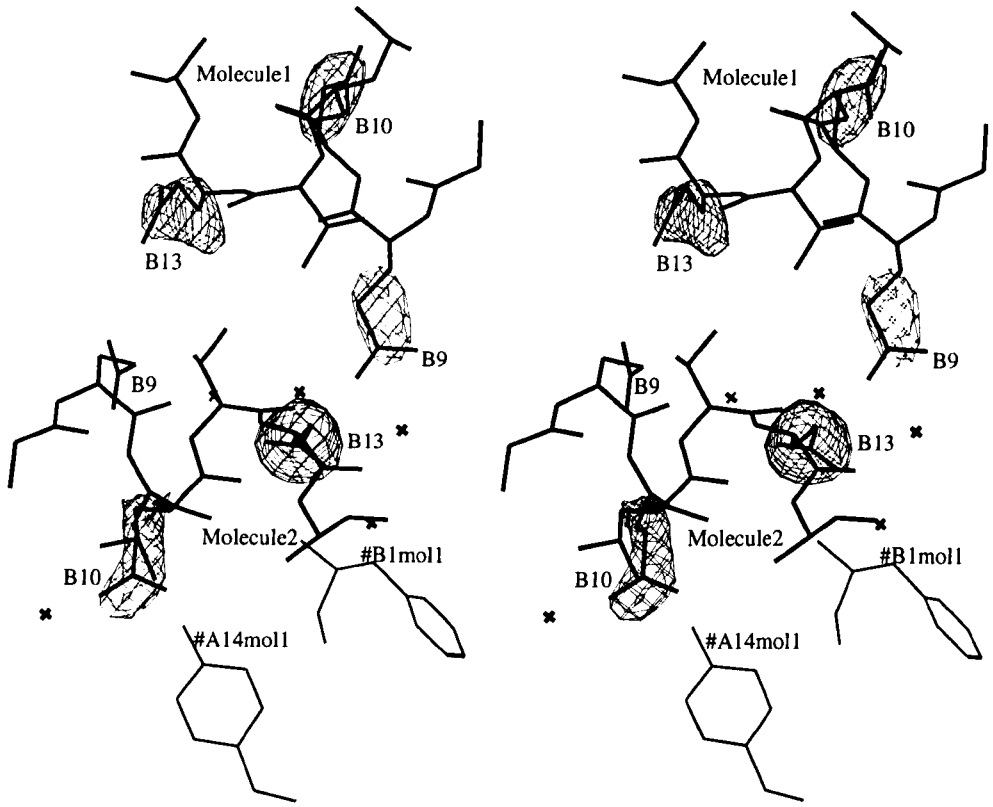

Figure 3.50: Final positions for the six central glutamic acids in A21G/B9E/B10E insulin; stereo view with $F_{obs}-F_{calc}$ electron density at $+3\sigma$. Symmetry-related atoms in thin lines.

hydrogen bonds with protein: B10 (molecule 2) comes into close contact with A14 Tyr Oη of a symmetry-related molecule 1, and B13 (molecule 2) makes a hydrogen bond with the N-terminal nitrogen of a symmetry-related B chain (molecule 1). Two glutamates, namely B9 (molecule 2) and B13 (molecule 1), seem to make no contacts at all, while the other two only make contacts with solvent.

At the resolution of the data (2.8Å), refinement is not totally reliable, and the temperature factors are essentially meaningless.

3.6.7 A4 Glu→Gln, B25 Phe→Tyr, des B30 single chain human insulin

B25Y(B29-A1)A4Q insulin crystallised in space group $P4_232$ with one insulin molecule in the asymmetric unit, and six insulin monomers grouped around the special positions with 32 symmetry to form four hexamers in the unit cell. The packing of the mutant (see figure 3.51) is equivalent to that seen for rat insulin II and insulin from the snake *Zaocys dhumnades dhumnades Cantor*, which has been reported by Wood *et al.*(1978) and Liang *et al.*(1984), respectively.

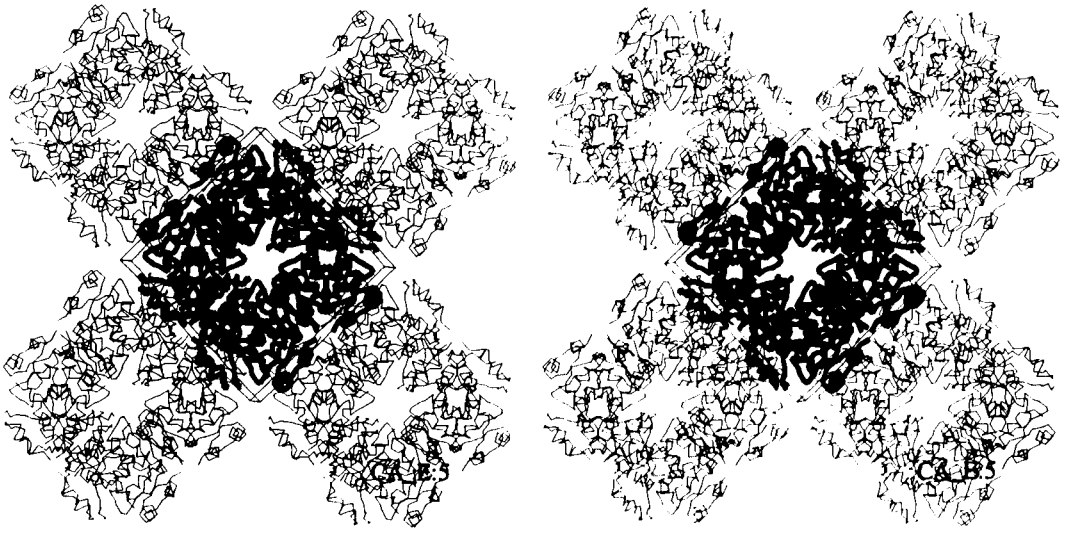

Figure 3.51: The packing of mutant B25Y(B29-A1)A4Q in the xz-plane. Stereo view of the C_α-atoms of 20 hexamers; cell edges also indicated.

Although the structure was not fully refined, some of the modifications were clearly interpretable from the electron density. The cross-link between residues B29 and A1, and the deletion of residue B30, is shown in figure 3.52. The conformation of residue B29 is such that after the β-sheet of residues B23 to B28 the chain immediately turns into the α-helix of what was the beginning of the A chain. Thus this α-helix is slightly longer in the mutant than in native insulin.

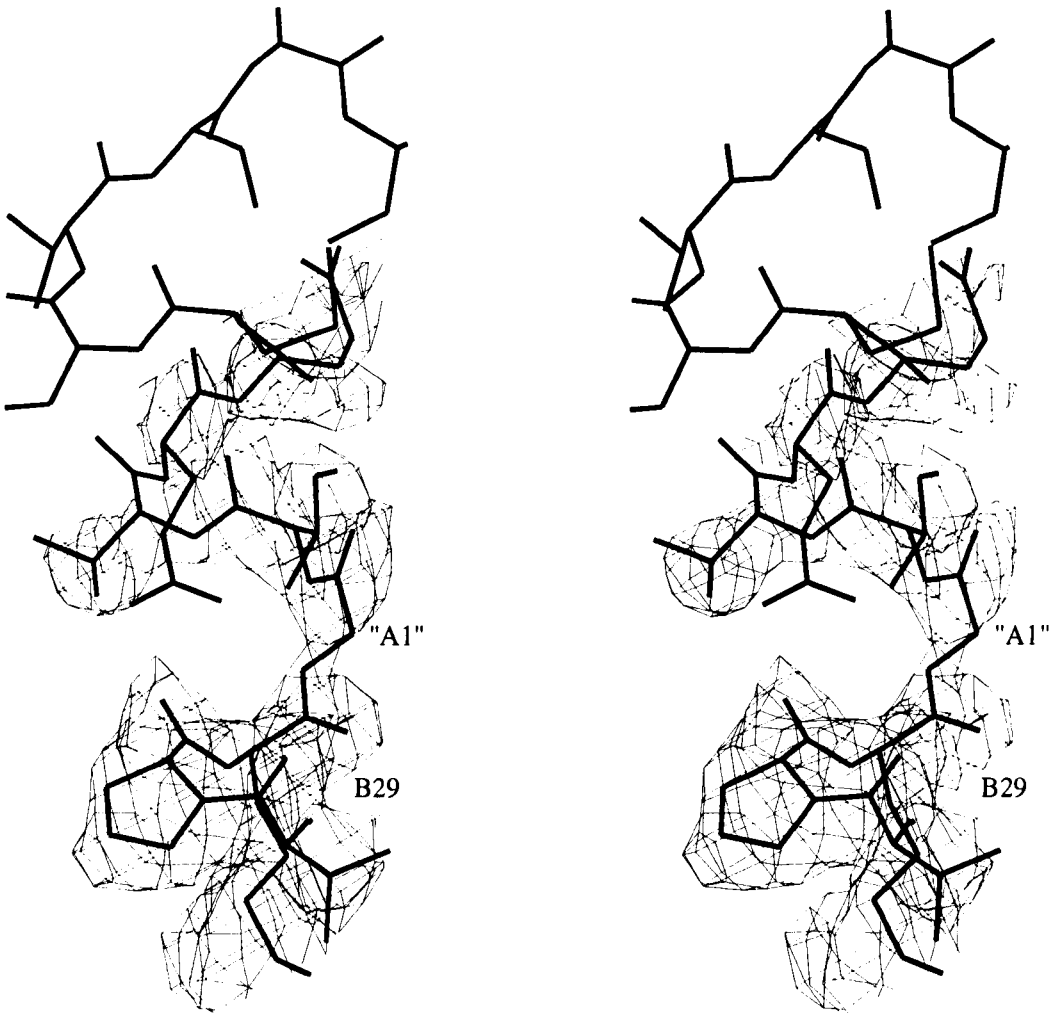

Figure 3.52: The peptide bond between B29 and A1 in B25Y(B29-A1)A4Q insulin. Stereo view with $2F_{obs}-F_{calc}$ electron density at $+1\sigma$ showing the elongated stretch of helix.

The conformation of the side chain of residue B25 from the native model was inappropriate for the mutant, clashing with the same residue from the twofold related monomer. A new position for this side chain was obvious from omit maps (calculated as described in section 3.5.1), even at limited resolution. Although the mutation from Phe to Tyr could not realistically be modelled, the new positions for the B25 residues result in convincing stacking of the aromatic rings (see figure 3.53). The amide group of the A4 Gln is poorly defined in the map. Although the change in charge must have some effect on the environment around this residue, the low resolution of the maps (2.8Å) prevented interpretation.

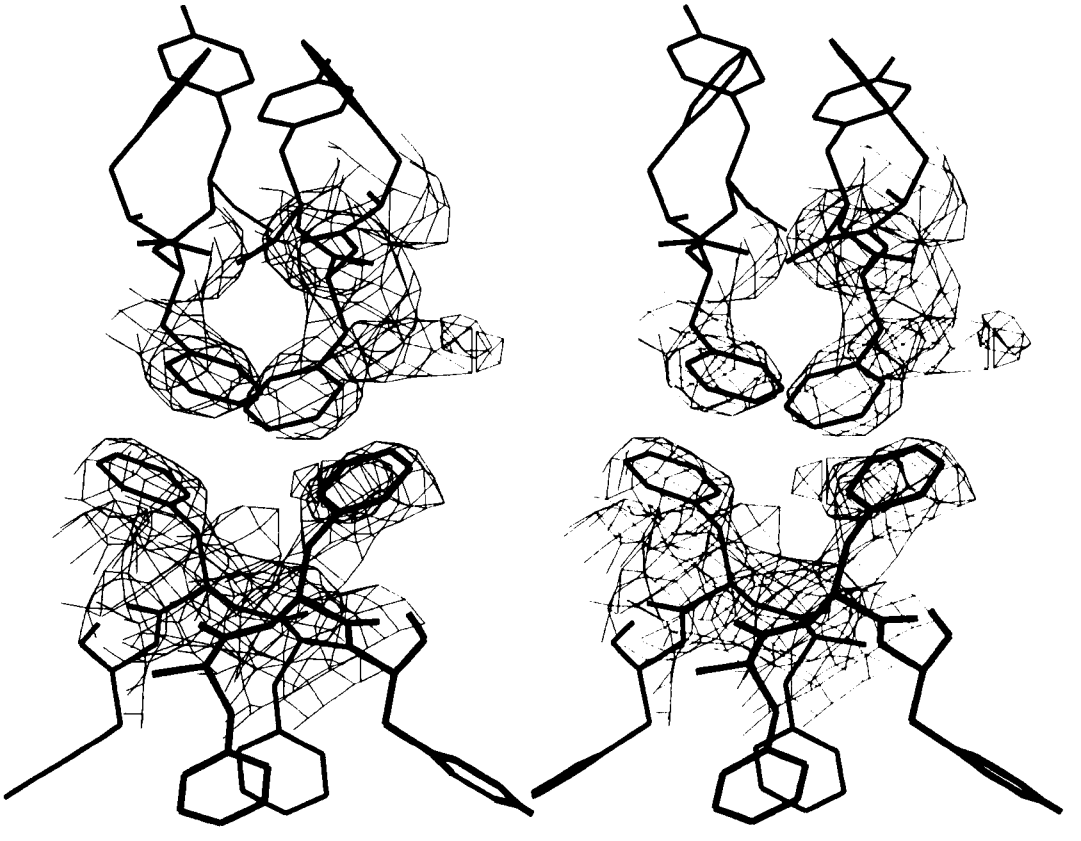

Figure 3.53: Aromatic stacking of twofold related B25 residues in B25Y(B29-A1)A4Q insulin. Stereo view with $2F_{obs}-F_{calc}$ electron density at $+1\sigma$.

The validity of the molecular replacement result was confirmed by the presence of convincing electron density for a zinc ion on the threefold axis at the point where three B10 histidine residues come together, as shown in figure 3.54.

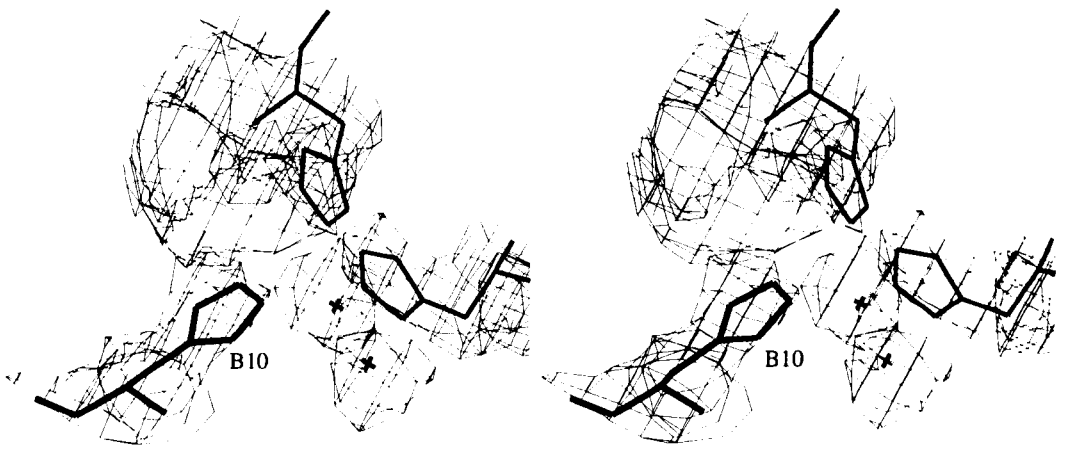

Figure 3.54: Electron density for zinc in B25Y(B29-A1)A4Q insulin. Stereo view of residue B10 (thick lines) and two symmetry-related B10 residues (thinner lines) with $2F_{obs}-F_{calc}$ electron density at $+1\sigma$. Position for zinc and a fourth ligand indicated by crosses.

Chapter 4

Insulin conformation

4.1 Introduction

The first insulin crystal structure described was that of the 2Zn insulin (Adams *et al.*, 1969). By that time, the ability of insulin to coordinate (for a description of the distinction between zinc *coordination* and zinc *binding*, see chapter 6 and Emdin *et al.* (1980)) more than two zinc positions had been recognised (Schlichtkrull, 1958), and the name 4Zn was devised to describe a second rhombohedral form thought to contain four zinc ions per hexamer (Bentley *et al.*, 1976). When the structure of human insulin in this form was solved and refined, it turned out there were five sites for zinc coordination, with a total stoichiometry of 2.67 zinc ions per insulin hexamer (Smith *et al.*, 1984).

In 1989, Kaarsholm *et al.* proposed the use of the terms T (tense) and R (relaxed), by analogy to the situation in haemoglobin, to describe the conformation of the structures rather than the zinc content. In what was called 2Zn insulin, the six N-terminal residues of the B chains are all extended, which was considered to be the more tense form, and the hexamer was therefore renamed 'T6 insulin'. 4Zn insulin was renamed 'T3R3 insulin', because in three of the monomers the N-terminus of the B chain is helical, extending the existing α-helix at residues B9 to B19. The helical

form is the more relaxed and is therefore named R. The monoclinic crystal form of insulin with phenol was assigned the name 'R6 insulin' because all six N-termini of the B chains in the hexamer are helical (Derewenda *et al.*, 1989). Kaarsholm *et al.* (1989) also suggest using T3T3' for 2Zn insulin to make clear the deviation from twofold symmetry in the hexamer.

In the early days of insulin structure determination, the groups from Oxford and China were using different numberings for the monomers within the asymmetric dimer of T6 insulin. At some stage it was decided to adhere to the Chinese numbering convention, where the monomer that does not change conformation upon addition of salt, considered to be the more conserved monomer, was designated the name 'molecule 1'.

The differences at the N-terminus of the B chains result in hexamers with quite different properties. In the T6 hexamer, the zinc ions coordinated by the B10 His residues are almost on the surface of the hexamer, in a shallow depression, from which they can easily diffuse out of the molecule. In the R6 hexamer, on the other hand, three α-helices come together in this depression, effectively shielding the zinc ions from the environment. This property is illustrated in figure 4.1, in which the T and the R 'faces' of the T3R3 hexamer are shown, representing the T and R character in one structure.

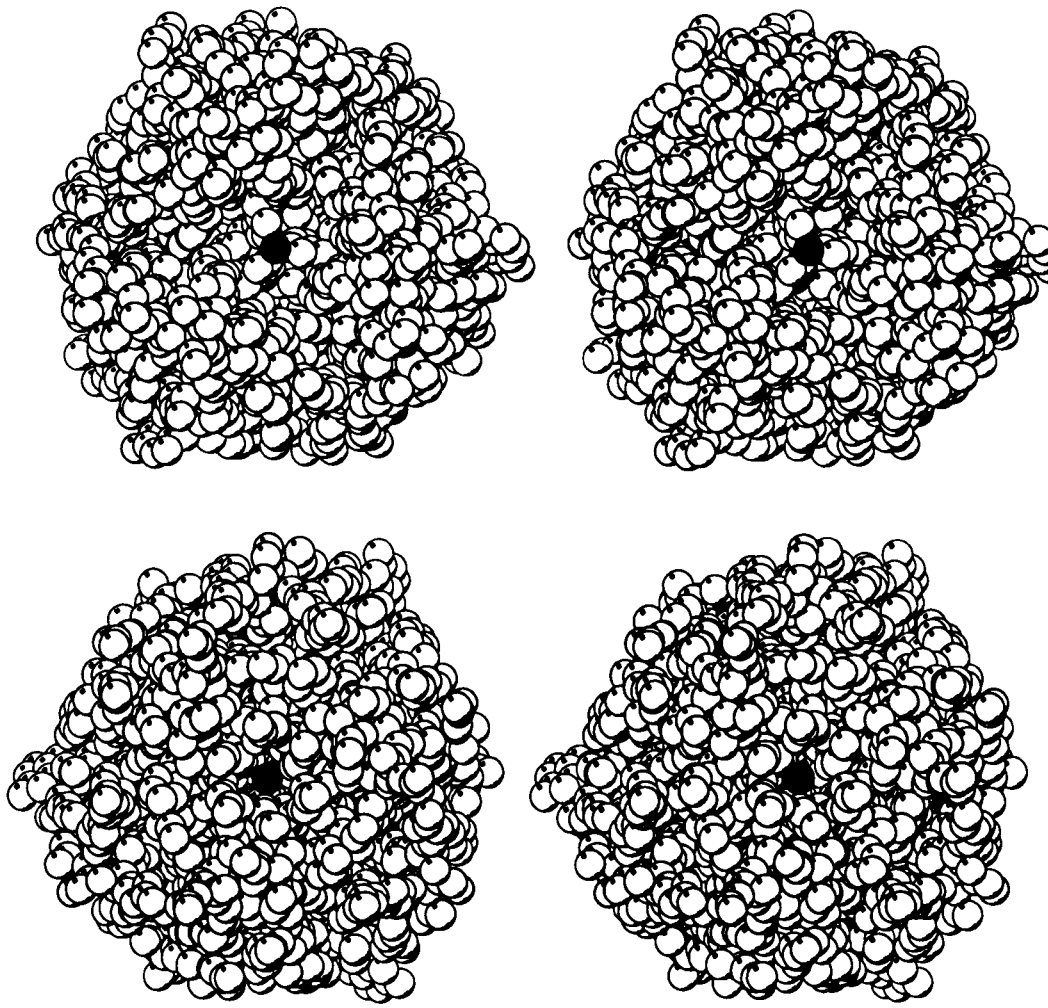

Figure 4.1: Stereo views of the T and R faces of the T3R3 hexamer. Produced with MOLSCRIPT (Kraulis, 1991).
a) T face, showing exposed zinc ion in the center in black;
b) R face, showing zinc ion in black at the end of a narrow channel formed by three α-helices.

From the observation that the zinc ions in the R6 hexamer are unable to diffuse in and out of the molecule, it was concluded that this might be the most stable hexamer species (Kaarsholm *et al.*, 1989). It has been shown that the R6 hexamer is stable in solution in the presence of phenol (Jacoby *et al.*, 1996). The most commonly observed monomer species in solution is a T-conformation monomer. Unfortunately, the molecule is either genetically engineered to stay monomeric in solution (Ludvigsen *et al.*, 1994) or the solution itself is manipulated in some way (*e.g.* very low concentration of insulin), which makes it difficult to determine the conformation of the native monomer in its active form. Ciszak and Smith (1994) have postulated that the insulin hexamer probably has to change to T6-conformation before the zinc ions can be released and the hexamer can dissociate into dimers and monomers.

The conformations of the N-termini of the B chains have an influence on the coordination of the zinc ions: in a T3 trimer the zinc coordination is octahedral, whereas in an R3 trimer it is tetrahedral. This is because in the R-conformation the B6 Leu side chain comes so close to the B10 His that there is no room for off-axial water molecules in an octahedral coordination (Smith *et al.*, 1984).

4.2 A different hexamer, T3R3^f

In 1995 Smith and coworkers (Ciszak *et al.*, 1995) announced the discovery of a new type of insulin hexamer, in which the first three residues of the B chain of molecule 2 are in extended conformation in a T3R3-like hexamer. They named it the 'T3R3^f' hexamer, where 'f' stands for 'frayed'. On re-examination it emerged that even the structure they determined in 1984 (Smith *et al.*, 1984) and the structure first described as 4Zn insulin (Bentley *et al.*, 1976) had this conformation.

Independently of the structure determinations by Smith and coworkers, the supposedly T3R3 structures of B13Q, B9H and CoI were all determined to have

the first three residues of the B chain of molecule 2 extended, resulting in a T3R3^f hexamer rather than a T3R3 hexamer. The presence of surplus electron density beyond the side chain of residue B3 Asn, the lack of electron density for residue B1 Phe and the observation that the Ramachandran angles for residues B2 Val and B3 Asn were in unfavourable regions of the plot, pointed in the direction of a necessary change of conformation for those residues. The B3 Asn side chain in the new position continues the hydrogen bonding pattern of the α-helix, and the Ramachandran angles of the residues in the new conformation are in the most favourable regions. The refined conformations of the first three residues of the modified insulins are shown in figure 4.2, along with the helical conformation of the T3R3 hexamer.

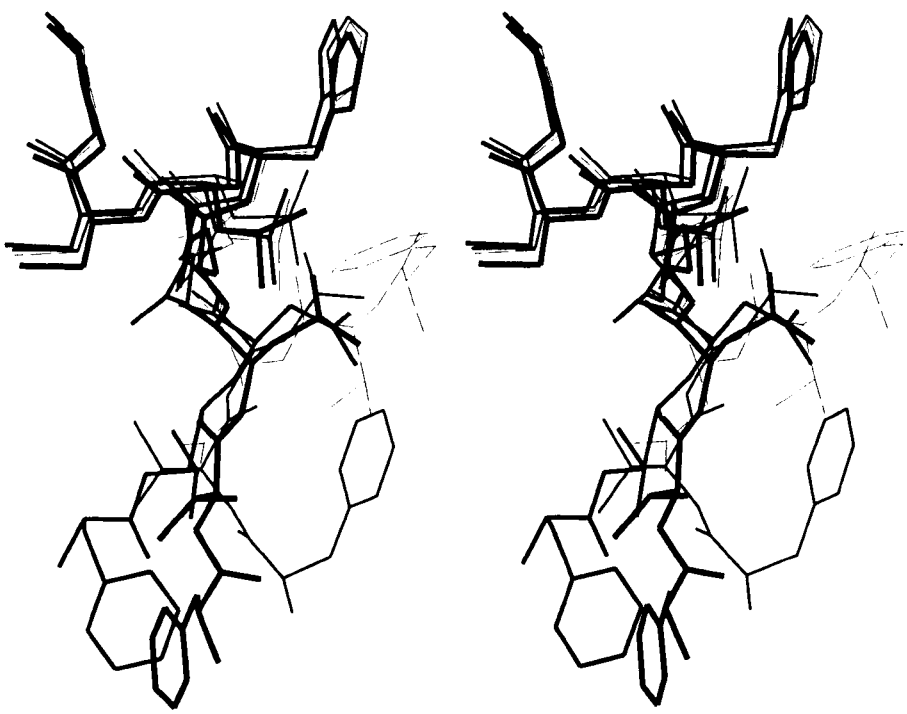

Figure 4.2: Stereo view of the N-terminus of the B chains of T3R3 insulin and three modified insulins in T3R3^f-conformation. Thin lines: CoI insulin; medium thick lines: B9H insulin; thick lines: B13Q insulin. For comparison: T3R3 insulin in dashed lines.

Upon careful examination of electron density maps of native T3R3 insulin it is deemed highly likely that this structure also exhibits the T3R3^f-conformation. This will be investigated by further refinement.

4.3 The T → R transition in modified insulin

The structures of more than one crystal form, and more than one hexamer conformation, have been determined for some of the modified insulins described in this thesis. B9H insulin forms T3R3^f and R6 hexamers in accordance with the crystallisation protocols (high salt and phenol conditions, respectively). Of the B13Q mutant, three crystal forms are now known: zinc-free T3R3, 2Zn T6, and 4Zn T3R3^f. The first two are described in Bentley *et al.* (1992), the latter in this thesis. The conformation of the zinc-containing T3R3^f structures is largely as expected, albeit the zinc content is different from that of the native isomorph (see chapter 6). No solution study data are available for this form. The behaviour of the mutant is unexpected, both in solution and in the crystal. The metal-free form in the crystal has T3R3-conformation, indicating either that the B13 Gln side chain itself or one of the additives in the crystallisation solution favour R- over T-conformation. The zinc-containing form produces T6 crystals, isomorphous to native insulin. In solution, however, the metal-free and metal-containing forms have the opposite behaviour: T6 for the metal-free and T3R3 for the metal-containing hexamer (Wollmer *et al.* (1989) and Bloom *et al.* (1995), respectively). The metal-containing form is present in the T3R3-conformation without addition of halides. Halides in the absence of metal do not affect the conformation, since they normally have their effect through the metal ion (Bentley *et al.*, 1992). Solution studies on the influence of phenol on the mutant show the presence of the typical R6-conformation (Wollmer *et al.*, 1987). The most exciting finding from solution studies is that the mutant can form R6 hexamers without phenol, in the presence of thiocyanate ions (Brzović *et al.*, 1994), thus by-

passing all the rules about the Hofmeister series and phenol (de Graaff *et al.*, 1981).

The B8S/B13Q/B30amide mutant has been described in T3R3 and R6 crystal forms (Dodson *et al.* (1991) and this thesis, respectively). The R6 crystal was grown following the monoclinic protocol, and has the expected conformation. The low-salt zinc protocol resulted in crystals with distorted T3R3 hexamers. The conformation of the N-terminus of the B chain in the T3 trimer is unusual: the residues have a completely different position from that seen in the native T3R3 hexamer. This is due to the B8 Ser residue having to adopt an allowable chirality in a position taken up by a glycine residue in right-handed conformation in the native structure.

Although CoI insulin has, as yet, only been described in one crystal form, this is unusual in itself. The crystallisation conditions would yield R6 hexamer crystals with zinc ions, but yielded T3R3 hexamer crystals in the case of cobalt ions. The standard phenol concentration of 0.67% is not high enough to drive Co^{2+} into a tetrahedral conformation in both trimers, necessary for the total T $\rightarrow$ R structural transition.

4.4 Six glutamines in a single and a triple mutant

Both B13Q and B8S/B13Q/B30amide mutants were designed to have reduced electrostatic repulsion in the core of the hexamer. Markussen *et al.* (1987) proposed a model for the six central glutamine residues, regularly arranged in a 'ring'-structure with well-defined hydrogen bonding (see figure 3.2). However, in neither of the mutants in any of the forms known, do the glutamines exhibit this regularity (Bentley *et al.*, 1992, Dodson *et al.*, 1991). Moreover, even in hexamers of the same conformation, namely the T3R3 forms of B13Q with zinc and high salt concentrations, zinc-free B13Q, and low-salt B8S/B13Q/B30amide, the glutamine residues take up varying positions (see figure 4.3). As can be seen from the figure, the B13 residues in molecule 2 of the B13Q $T3R3^f$ structure (grown in the presence of zinc and high salt

concentrations) has dual conformation, one of which overlaps reasonably well with the same residue in the other structures. The B13 residues in the T3 trimer, however, take up different side chain positions in all three structures, getting involved with different hydrogen bonding.

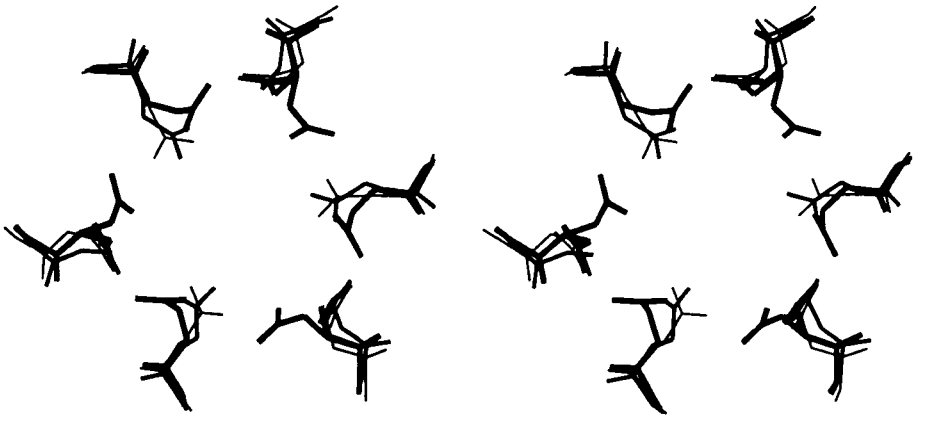

Figure 4.3: Stereo view of the B13 glutamine residues in three T3R3 structures of modified insulins. Thickest lines: B13Q T3R3^f structure; medium thickness lines: zinc-free B13Q structure; thin lines: low-salt zinc B8S/B13Q/B30amide structure.

4.5 Summary

The conformation of modified insulins has been rationalised in comparisons with previously known native and mutant insulin in various crystal forms. Seemingly subtle changes in either the modified insulin itself (*e.g.*B13Q insulin where one residue is replaced by one of exactly the same size) or the crystallisation conditions (*e.g.*Co^{2+} instead of Zn^{2+} in CoI insulin) have led to unexpected conformations for some of the modified insulins. Independently of Smith and co-workers (Ciszak *et al.*, 1995) the T3R3^f conformation has been established for all modified insulins isomorphous with native 4Zn insulin.

Chapter 5

Aggregation of insulin

5.1 Introduction

The assembly of insulin into hexamers is well understood. In the B-cells of the Islets of Langerhans single chain preproinsulin is converted to proinsulin, in which the A and B segments fold spontaneously into the insulin conformation. It then forms dimers, and with the zinc ions present in the B-cells hexamerisation takes place, after which conversion to mature insulin hexamers occurs. These hexamers then precipitate in crystalline form owing to the pH (around 5.5) and the presence of zinc and calcium ions. Upon extracellular release, the hexamers dissociate into dimers and eventually into monomers, which are the biologically active species.

The crystallisation of insulin as occurring in nature, can be mimicked in the laboratory. Several crystallisation protocols have been successful, yielding various aggregation states and different conformations (see also chapter 4). The assembly of proteins into polymers is stabilised by salt bridges, hydrogen bonds, dispersion forces and by the entropic benefit associated with water released by the burial of non-polar surfaces. An overview of the aggregation of insulin is shown in figure 5.1.

The tendency of insulin to form aggregates causes problems in the diabetic patient, because of the slow release of insulin from the injection site. Genetically

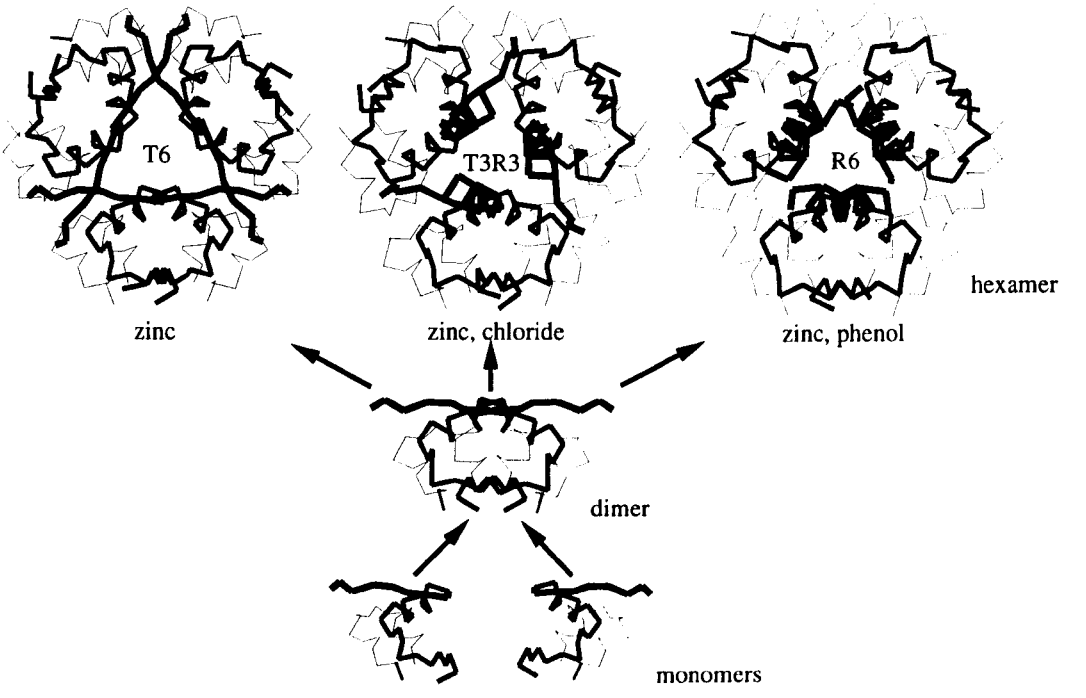

Figure 5.1: Aggregation of insulin in various conformations.
A chains in thin lines, B chains in thicker lines, with N-termini in thickest lines. Monomers and dimer in T-conformation. Upon hexamer formation the B chain N-terminus may become helical, resulting in R-conformation.

engineered insulins are therefore being produced which are not aggregated after injection and therefore quickly released into the blood stream. On the other hand, the production of more stable hexamers has led to slower acting insulins which allow for a more constant basal level of insulin in diabetics. Knowledge about the dimer and hexamer forming surfaces was used in the design of the mutants.

5.2 Insulin dimerisation

The dimer forming surface of the insulin monomer is almost planar, with mainly non-polar residues, as shown in figure 5.2. It comprises the B chain C-terminal

residues B22 to B30, thereby including residues B24 to B26 which are in β-strand conformation and form an antiparallel β-sheet in the dimer. The B chain α-helix lies against the monomer surface, with residues 9, 12, 13 and 16 pointing outwards. In the dimer, the two B13 glutamic acid residues are brought together, causing electrostatic repulsion. In order for dimerisation to take place, this electrostatic repulsion must be overcome by surplus favourable energies from other interactions. Residues B20 and B21 of the loop between the B chain α-helix and β-strand, also lie on the dimer forming surface.

Figure 5.2: Stereo view of the dimer forming surface of the insulin monomer, on the right-hand-side in this view. Produced with MOLSCRIPT (Kraulis, 1991).

The shape of the dimers as seen in various crystal structures is slightly different, mainly due to the difference in conformation of the N-terminus of the B chains. The dimers of the four native insulin species relevant to this thesis are shown in figures 5.3, 5.4, 5.5 and 5.6.

Figure 5.3: The native orthorhombic insulin dimer, T2-state. View of the smallest dimension on the left; view of the largest dimension on the right (perpendicular to the first view).

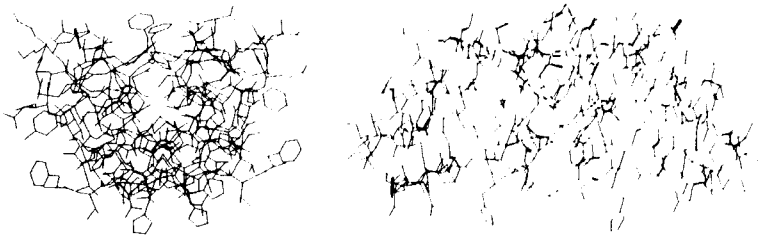

Figure 5.4: The dimer from native insulin in rhombohedral form, T6-state.
View of the smallest dimension on the left; view of the largest dimension on the right (perpendicular to the first view).

Figure 5.5: The dimer from native insulin in rhombohedral form, T3R3-state.
View of the smallest dimension on the left; view of the largest dimension on the right (perpendicular to the first view).

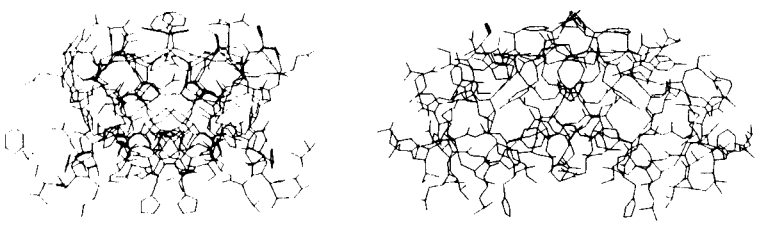

Figure 5.6: The dimer from native insulin in monoclinic form, R6-state.
View of the smallest dimension on the left; view of the largest dimension on the right (perpendicular to the first view).

5.2.1 Dimer stability

The insulin dimer is stable in aqueous solution between pH 2 and 8. The packing interactions that stabilise the insulin dimer are adaptable, even where they involve close and apparently complementary contacts. Therefore, non-polar mutations of the monomer, such as B12 Val→Leu (Dodson *et al.*, 1993), were ineffective in reducing the association of monomers into dimers. Moreover, the residues in the β-strand of the insulin monomer, up to B25, are thought to be important in receptor binding, and mutation of those residues might lower the potency of the mutant as a therapeutic agent. The more important factor in insulin dimerisation seems to be the electrostatic interactions. Introduction of more negative charges in the region of residues B9 to B13 reduces dimerisation in solution. The mutation of B9 to Asp produces one of the most effective monomers at neutral pH (Brange *et al.*, 1988). Glutamic acid is more flexible than aspartic acid, and thus more capable of moving away from the site of repulsion.

The B9D/B27E and A21G/B9E/B10E mutants are indeed found to be monomeric under physiological conditions (Brange *et al.*, 1988, Xiao, 1990). The crystallisation environments (in particular the pH), however, can be such that dimerisation (or even hexamerisation) can still take place. In the case of B9D/B27E insulin, the carboxylates of B9 and B13 seem partially protonated and can therefore form hydrogen bonds with each other. There are also hydrogen bonds to water molecules, for which the water molecules could be the hydrogen donors. The degree of protonation is unclear, because the pH of the crystallisation is unknown. In the case of A21G/B9E/B10E insulin, for which the crystals were grown at neutral pH, protonation of the glutamates is unlikely. Nevertheless, this mutant forms dimers in the crystal, by virtue of spatial separation of the six glutamates, and hydrogen bond formation with other parts of the protein and water molecules.

The dimer species of insulin is virtually the same, apart from the conformation of the N-terminus of the B chain, throughout the whole range of aggregates and crystal forms (see figures 5.3 to 5.6). Although the packing of the A21G/B9E/B10E mutant is quite different from that of the native and B9D/B27E insulins, this is only due to a rearrangement of the flexible C-termini of the B chains. As viewed along the y-direction, which is along a 2_1 screw axis, it can be seen from figure 5.7 that the A21G/B9E/B10E molecule is merely rotated with respect to the native molecule. This, and the changes at the C-termini of the B chains, allows for the different packing in a slightly smaller volume.

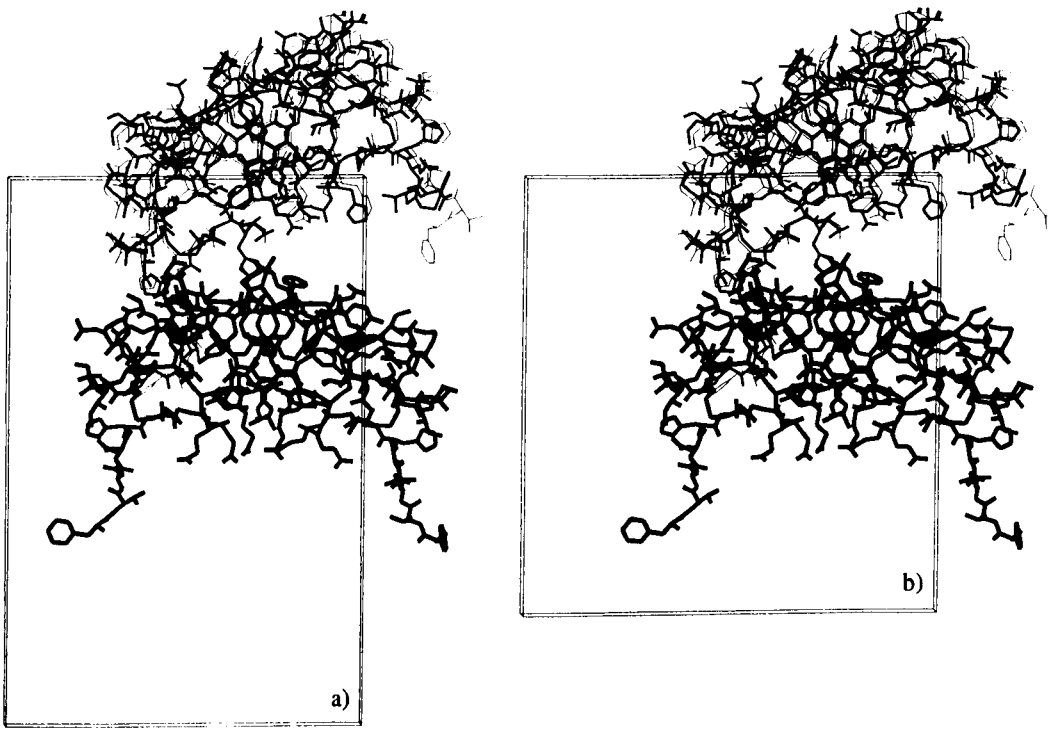

Figure 5.7: Three orthorhombic dimers. Native and B9D/B27E insulin (thin lines) are isomorphs, and A21G/B9E/B10E insulin (thick lines) has different packing.
a) View of the three dimers along y, showing the cell of native insulin;
b) The same view, showing the cell of A21G/B9E/B10E insulin.

5.3 Insulin hexamerisation

The insulin dimer is wedge-shaped (see for example figure 5.3), rounded like an orange-segment. The hexamer forming surfaces are loosely planar, along the long dimension of the wedge, almost 70° apart when looking at the triangular side of the wedge. The differences at the N-termini of the B chains cause the contacts between dimers to be quite different in the various insulin hexamer types. In the T6 hexamer of native 2Zn insulin, only around half of the long dimension of the wedge is involved in dimer-dimer contacts, whereas in the R6 hexamer (the native monoclinic form) almost all of the residues along the long side of the wedge have contacts with other dimers. All hexamers have residues B9, B10 and B13 in the core.

In the T6 hexamer there is strong hydrogen bonding between A14 tyrosine residues from both molecules, to that of a symmetry relative from the other molecule. The B13 glutamic acid groups form hydrogen bonded pairs but usually do not come into close contact in the centre. The extended N-termini of both B chains make close contacts with the A chains of symmetry-related molecules, with the B1 Phe side chains in a hydrophobic pocket between the main A chain and A14 Tyr of its neighbour. There are intricate contacts between residues B9 Ser and B10 His: from B9 of molecule 1 to B10 of a symmetry-related molecule 1 and from B10 of molecule 1 to B9 of the other symmetry-related molecule 1, and *vice versa* for the B9 and B10 of molecule 2. In the R6 hexamer the A14 tyrosine residues do not seem to be involved in hexamer formation; some can not be seen in electron density maps. Also, the contacts between residues B9 and B10 are irregular. Some of the B13 glutamate groups come into close contact with each other. The N-terminal stretches of the B chain α-helices come close together to form a channel of approximately 8Å diameter, shielding the zinc sites, with contacts to the A7-A11 loops. In the T3R3 hexamer, the A chain of molecule 1 is not involved in dimer-dimer contacts at all. The contacts between B9 and B10 residues are similar to those in the T6 hexamer, with both

conformations of residue B10 of molecule 2 making contact with B9 of a symmetry-related molecule 2. There is no close proximity of the B13 glutamic acid groups. The N-terminus of the B chain of molecule 1 makes contact with residues B16 Tyr to B21 Glu of a symmetry-related molecule 2. One of the strongest contacts seems to be that between B21 from molecule 2 with B3 of a symmetry-related molecule 1, which limits the flexibility of the B21 Glu side chain.

The insulin hexamer is torus-shaped with dimensions 35Å by 50Å. It essentially has 32 symmetry. In the T6-state hexamer in the rhombohedral crystal, this symmetry is nearly exact. The threefold axis lies along the crystallographic *c*-axis, and the differences between the two T-state monomers in the asymmetric unit are only minor. In the T3R3 hexamer in the same crystal form, the 32 symmetry breaks down. The threefold axis is exact, but the local twofold symmetry is distorted, due to the changes at the N-terminus of the B chain of molecule 2 (see figure 5.5). In the R6 monoclinic structure both the threefold and the twofold axes are not exact. The monomers, however, are similar, and the non-crystallographic symmetry is reasonably well preserved (see figures 3.7 and 3.6). In the case of the B25Y(B29-A1)A4Q mutant, the hexamer is situated on a special position of 32 symmetry in the cubic unit cell. The monomer is therefore the asymmetric unit.

In a solution free of metal ions, insulin exists as a mixture of monomer, dimer, tetramer, hexamer and higher aggregates, proportions depending on the concentration (Blundell *et al.*, 1972). In the presence of zinc ions, as in the B-cells of the Islets of Langerhans, the hexamer species prevails. It has been found that other divalent metal ions (*e.g.* Ni^{2+}, Co^{2+}, Cd^{2+} and Cu^{2+}) can play the same role in the hexamerisation of insulin (Schlichtkrull, 1958, Hill *et al.*, 1991). Zinc sites associated with a T-state molecule are seen to have octahedral coordination, whereas those associated with insulin in the R-state have tetrahedral coordination.

5.3.1 Hexamer stability

The insulin hexamer is stable in aqueous solution between pH 5 and 8. Because both polar and non-polar residues are buried between the dimers, the packing of dimers into a hexamer is much looser than that of monomers into a dimer. Six B13 glutamic acid residues come close together in the core of the hexamer, causing electrostatic repulsion which must be reduced in order for hexamerisation to take place. The coordination of zinc in the core balances the charges, and is the driving force in hexamer formation in insulin. The R6 hexamer seems to be the most stable hexamer species in solution. This is probably because of the favourable contacts of the N-termini of the B chains. Not only are more residues involved in dimer-dimer contacts in the R6 hexameric state, but the N-terminal stretch of helix closes off the solvent channel towards the zinc ions, disabling diffusion of zinc ions into the solvent. In the T6 hexamer in solution there is a constant exchange of zinc between protein and solvent (Smith *et al.*, 1984).

The stability of the hexamer may be enhanced through various measures, to improve prolonged-acting insulin in therapy. In the B13Q mutant the electrostatic repulsion of the B13 glutamate groups is removed by substitution with the neutral residue glutamine. The mutant readily forms hexamers under various crystallisation conditions, even without the presence of zinc or other divalent metal ions (Bentley *et al.*, 1992). Therefore, the absence of repelling negative charges in the core of the hexamer is enough to nullify the requirement of zinc in the hexamerisation of insulin. In B8S/B13Q/B30amide insulin the substitution of B13 Glu with Gln has the same effect as in the single mutant B13Q. The mutant crystallises readily as R6 hexamers in the presence of phenol. It has also been crystallised as T3R3 hexamers under low salt conditions, with the conformation of the T-state quite different from that in native insulin (Dodson *et al.*, 1991), see also chapter 4. In the B9H mutant, more binding sites for zinc have been introduced. In the rhombohedral form of this mutant the zinc coordination is well-defined, if geometrically unusual.

The mutant adapts to different crystallisation conditions by forming monoclinic R6 hexamers, creating a pool of zinc in the core of the hexamer, with some of the B13 glutamic acid residues taking part in zinc coordination. In that structure, all six glutamic acid residues may be fully ionised without causing repulsion in the core of the hexamer. The CoI insulin shows that Co^{2+} can indeed play the same role in insulin hexamerisation as Zn^{2+}, both in octahedral and tetrahedral coordination. The B25Y(B29-A1)A4Q mutant crystallises as a fully 32 symmetric hexamer, in which the B25 side chains have found a position that allows twofold symmetry of the dimer axis. For mutant B9D/B27E it is possible to form hexamers when Ca^{2+} ions are present in the crystallisation solution, balancing the charges of the B9 Asp residues in the core of the hexamer (Turkenburg, 1992).

5.4 Packing of insulin in crystals

In the T6 conformation, the departure from twofold symmetry in the essentially 32 symmetric hexamer is due to packing contacts. Those for the 2Zn native insulin in rhombohedral crystal form are described in Baker *et al.* (1988). The loop A8 to A10 protrudes from the hexamer into a gap created by the same stretch of residues in a different hexamer along the threefold axis. This would cause B5 histidine from molecule 1 in one hexamer to collide with B5 histidine from molecule 2 in a different hexamer. To avoid this, one of them is rotated to another, conformationally stable, position. The packing of three hexamers in the unit cell is shown in figure 5.8 on the right, viewed along the threefold hexamer axis. The two hexamers in thin lines are each at different levels of z, along the direction of the threefold axis.

The packing contacts for native monoclinic insulin (the R6 hexamer), 4Zn native insulin in rhombohedral crystal form (the T3R3 hexamer), and native orthorhombic insulin (the T2 dimer) have been described by Derewenda (1990). In the R6 hexamer, the differences in the number of contacts from one chain to the

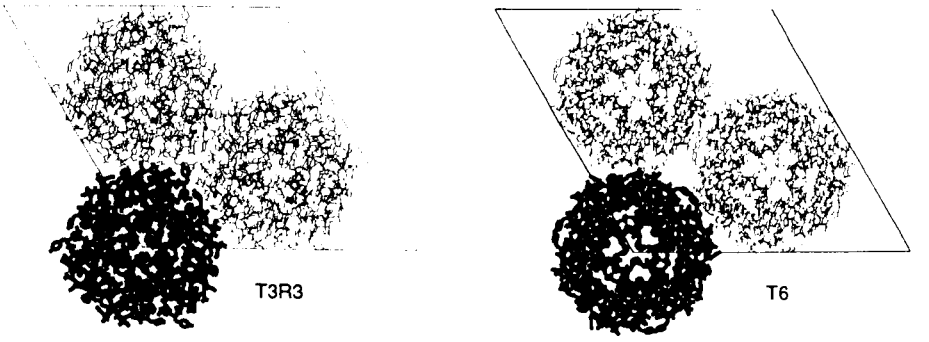

Figure 5.8: Packing of two rhombohedral crystal forms of insulin. T3R3-form ('4Zn') on the left, T6-form ('2Zn') on the right. Hexamers at three different levels of z around the 3_1 screw axis, in both cases.

next are marked, ranging from two to as much as ten per chain. Therefore, the symmetry within the hexamer is less exact. The packing of the R6 hexamer is indicated in figure 5.9, viewed along one hexamer's local threefold axis. It can be seen from this figure that the hexamers do not pack along this threefold axis; the symmetry-related hexamer in the unit cell is packing behind (in this view) the first hexamer, with its local threefold axis approximately perpendicular to the first. The packing in the plane perpendicular to the first hexamer's non-crystallographic threefold axis is shown to reveal the crystal contacts in that direction.

In the T3R3 hexamer, the asymmetry of the hexamer and the asymmetry of the crystal contacts are not attributable to the collision of the B5 histidines, as in the native T6 hexamer. This is due to a conformational change in the N-terminus of the B chain of molecule 2, which in itself constitutes a large deviation from exact symmetry. The crystal contacts of molecule 2 are therefore markedly different from that of molecule 1. The packing of three hexamers in the unit cell is shown in figure 5.8, alongside that for T6 hexamers in the same view.

In the orthorhombic dimer, B1 Phe of both molecules play an important role in the crystal contacts between the dimers. As can be seen from figure 5.3, these

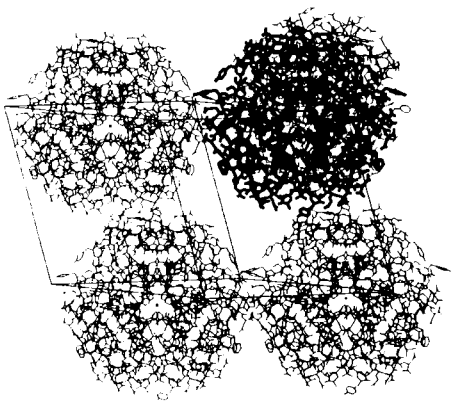

Figure 5.9: Packing of monoclinic insulin, a layer perpendicular to the local threefold axis of one of the hexamers in the unit cell.

residues protrude far from the bulk of the dimer. Residues A13 and A14, involved in hexamerisation in rhombohedral crystals, are involved in crystal contacts between dimers across crystallographic twofold axes in the orthorhombic crystal. The crystal contacts in orthorhombic insulin are fewer than in hexameric forms, reflecting higher solvent content. This is shown for one direction in figure 5.10.

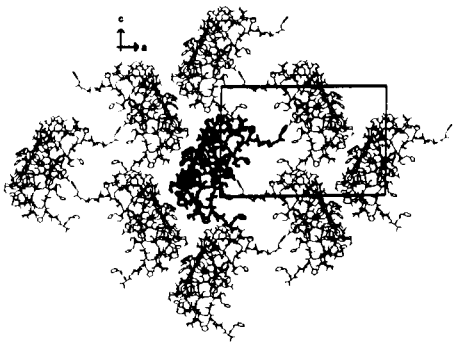

Figure 5.10: Packing of orthorhombic insulin. Direction compared to figure 5.7: rotated 90° along an axis perpendicular to the plane of the paper.

The packing of the A21G/B9E/B10E mutant is different from that of native orthorhombic insulin, which is shown in figure 5.10. The direction of this figure is

chosen to match that of figure 5.7, showing the different positioning of the native and mutant dimers in the cell. The packing of the mutant can be seen in figure 5.11.

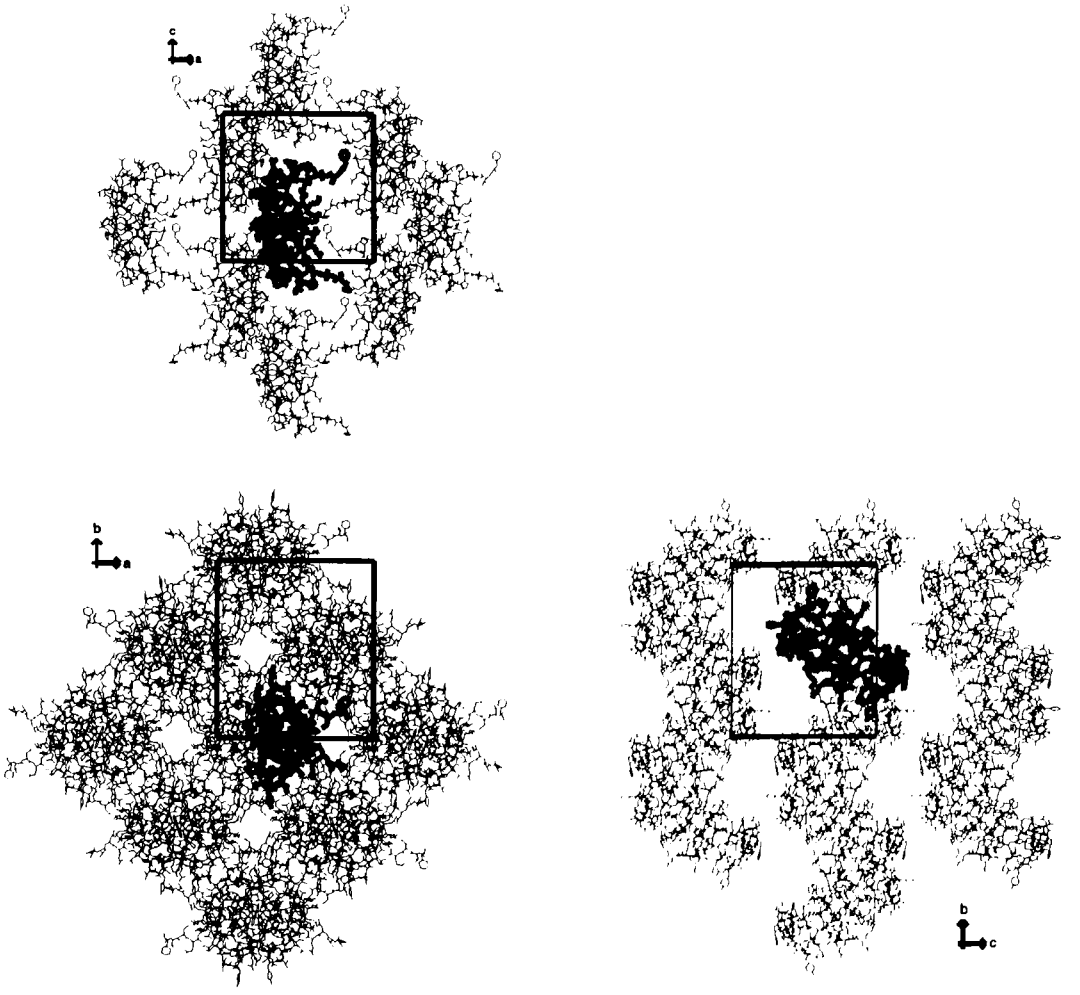

Figure 5.11: Packing of A21G/B9E/B10E insulin, shown along the three cell axes.

In comparing the top left part of this figure with figure 5.10, it is quite clear that the packing of the mutant in this plane is tighter than that of the native molecule, resulting in a smaller unit cell. The crystal contacts of the mutant are summarised in table 5.1. As can also be seen in figure 3.50, only one of the mutated residues is involved in crystal contacts, namely residue B10. This residue from both

molecules in the dimer is involved in contacts with an A14 tyrosine residue from the other molecule of a symmetry-related dimer. The new positions for the C-terminal residues in both B chains provide new crystal contacts, not observed in the native structure (Derewenda, 1990).

A chain contacts						B chain contacts					
chain	res†	mol‡	chain	res	mol	chain	res	mol	chain	res	mol
A	5	1	A	18	2	B	1	1	A	14	1
A	12	1	B	4	2	B	1	1	A	17	1
A	14	1	B	1	1	B	1	1	B	22	1
A	14	1	B	3	2	B	1	1	B	13	2
A	14	1	B	6	2	B	2	1	A	17	1
A	14	1	B	10	2	B	3	1	A	14	2
A	15	1	A	18	2	B	4	1	A	18	1
A	15	1	B	4	2	B	4	1	A	12	2
A	17	1	B	1	1	B	6	1	A	14	2
A	17	1	B	2	1	B	10	1	A	14	2
A	18	1	A	5	2	B	14	1	A	13	2
A	18	1	A	18	2	B	17	1	A	17	2
A	18	1	B	4	1	B	17	1	B	17	2
A	19	1	A	5	2	B	18	1	B	17	2
A	19	1	A	18	2	B	22	1	B	1	1
A	19	1	A	19	2	B	25	1	A	4	2
						B	27	1	B	30	2
A	4	2	B	25	1	B	1	2	A	14	2
A	4	2	B	25	2	B	1	2	A	17	2
A	5	2	A	18	1	B	3	2	A	14	1
A	5	2	A	19	1	B	4	2	A	12	1
A	12	2	B	4	1	B	4	2	A	15	1
A	13	2	B	14	1	B	6	2	A	14	1
A	14	2	B	3	1	B	10	2	A	14	1
A	14	2	B	6	1	B	13	2	B	1	1
A	14	2	B	10	1	B	17	2	B	17	1
A	14	2	B	1	2	B	17	2	B	18	1
A	17	2	B	17	1	B	25	2	A	4	2
A	17	2	B	1	2	B	30	2	B	27	1
A	18	2	A	5	1						
A	18	2	A	15	1						
A	18	2	A	18	1						
A	18	2	A	19	1						
A	19	2	A	19	1						

Table 5.1: Crystal contacts <3.6Å in A21G/B9E/B10E insulin. †res=residue; ‡mol=molecule.

B25Y(B29-A1)A4Q insulin makes few packing contacts, compared to other hexameric insulins. A detailed picture of the crystal contacts of the mutant is shown in figure 5.12, and they are summarised in table 5.2. The contact of residue 43 Tyr (for numbering of this mutant, see the legend to table 5.2) with its symmetry relative

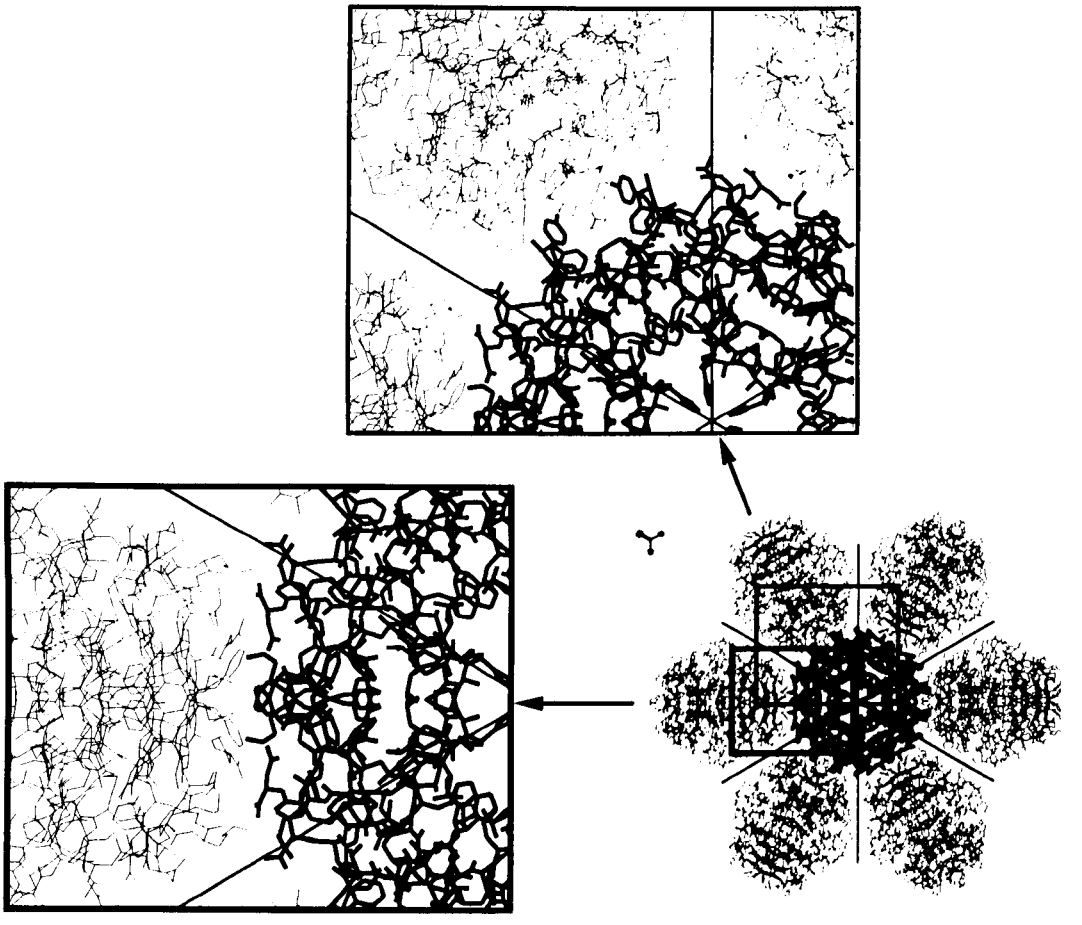

Figure 5.12: Detailed packing of B25Y(B29-A1)A4Q insulin. The two regions of close contact between hexamers in exploded views.

is included in the table because it constitutes an aromatic stacking interaction, although the distance is somewhat larger than the other contacts. The enlarged part on the left of figure 5.12 includes the contacts of 25, 29, 49 and 50, while the top half of the figure shows the contacts of residues 1, 22, 41, 42 and 43. Residue 5, the histidine residue which causes asymmetry in the native rhombohedral T6 hexamer, is surrounded by solvent in the mutant. Exact 32 symmetry is therefore possible. The solvent content of the mutant and the lack of precise contacts are probable causes of the limited diffraction.

from residue	to residue	approximate distance (Å)
1	22	2.6
1	43	3.4
25	25	3.4
29	49	2.7
29	50	2.7
41	43	3.3
42	42	3.4
43	43	3.9
50	50	3.4

Table 5.2: Crystal contacts <3.5Å in B25(B29-A1)A4Q insulin. The N-terminus of the mutant is equivalent to residue B1 in native insulin, and the residues are numbered consecutively, from 1 to 50.

A stereo view of the contents of one unit cell of B25Y(B29-A1)A4Q insulin is shown in figure 5.13, with the body diagonals of the cube indicated from the appropriate corner of the cell to the midpoint. The hexamers are located on points of 32 symmetry, $\frac{1}{4}$ along each body diagonal (from different corners), in a tetrahedral arrangement with respect to each other.

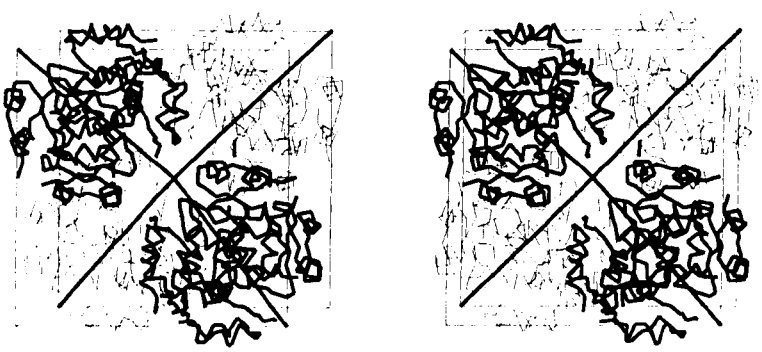

Figure 5.13: Stereo view of packing of B25Y(B29-A1)A4Q insulin within one unit cell, showing the tetrahedral arrangement of four hexamers, with the appropriate half of each of the body diagonals indicated.

The three-dimensional packing of the mutant is shown in figure 5.14, viewed along the threefold axis of the central hexamer. The four tetrahedrally arranged

hexamers (see figure 5.13 in different view) are shown in thick lines in the main overlay. The three non-central ones in the tetrahedrally arranged set of four (shown separately in the circle to the right of the middle overlay) are behind the layer of hexamers shown in the circle to the left of the middle overlay. These tetrahedrally related hexamers do not make any contacts shorter than 3.5Å with the central hexamer.

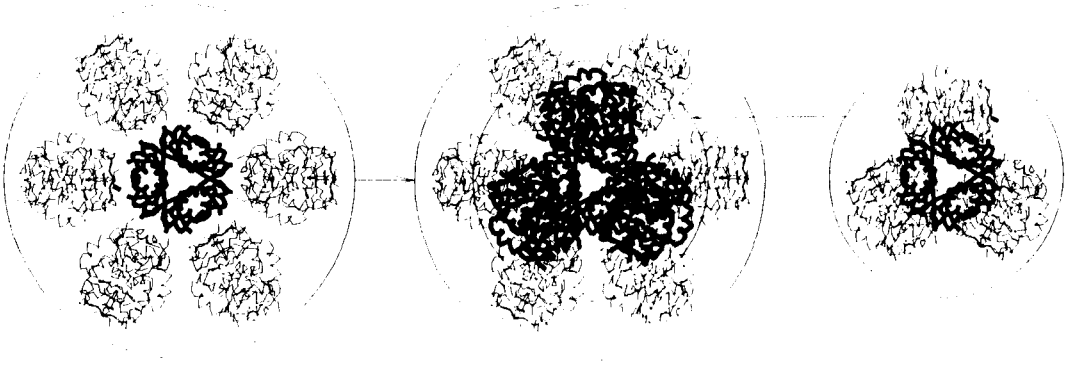

Figure 5.14: Extended packing of B25Y(B29-A1)A4Q insulin. Components of the middle diagram indicated: contents of one unit cell (shown on the right) superposed on one full layer of hexamers (shown on the left).

The crystal contacts of the isomorphous modified insulins B13Q, CoI, B9H and B8S/B13Q/B30amide are much the same as those of the native equivalents. The orthorhombic mutant B9D/B27E is quite different from native orthorhombic insulin at the N-terminus of the B chain of molecule 1, and therefore has different crystal contacts at that point. Residue B2 of molecule 1 has an additional contact with residue A14 of molecule 2 of another dimer, and residue B1 of molecule 1 is in contact with residue A2 of molecule 2 of yet another dimer.

5.5 Summary

Aggregation and packing variations have been rationalised using existing structural knowledge. Mutant studies have drawn attention to the possibility of specific choice of aggregation state of insulin for therapy (mutations may increase or reduce aggregation specifically). The crystallisation environment does, however, not always represent a suitable environment for the desired aggregation state, resulting in for example dimers in the crystal form for mutants which are monomeric under physiological conditions.

Chapter 6

Metal binding in insulin

6.1 Introduction

Insulin storage vesicles in humans and many other species contain high concentrations of Zn^{2+} and Ca^{2+} ions (Howell *et al.*, 1975). Zinc plays an important role in insulin hexamerisation, which is closely related to some of the processes in insulin biosynthesis and storage (Emdin *et al.*, 1980). Sudmeier *et al.* (1981) established that the 2Zn insulin hexamer is also capable of binding calcium. Detection of calcium in zinc insulin crystals by X-ray diffraction is difficult since calcium and water only have a slightly different number of electrons (Hill *et al.*, 1991). A zinc-free insulin hexamer crystal can be artificially produced by soaking out the zinc from an existing 2Zn crystal with EDTA, after which the B10 histidine residues normally involved in zinc coordination are in the same conformation, indicating they are held in place by the hexamer as a whole (Hill *et al.*, 1991).

Insulin has two sites for zinc binding seen in all zinc containing insulin hexamers, on the (local) threefold axis, about 15Å apart on either side of the centre of the hexamer. Other divalent metal ions can be substituted easily for Zn^{2+} in these positions, a phenomenon investigated first by Schlichtkrull (1958). In crystals of native insulin grown with high salt concentrations, three more, off-axial, sites for

zinc binding occur, with B5 His residues in molecule 2 taking part in the coordination of the metal. The B10 His residues in molecule 2 take up two conformations, one coordinating to the zinc ion on the threefold axis, the other coordinating to the off-axial zinc ion. The anion of the salt acts as additional ligand for the metal.

Emdin *et al.*(1980) make a distinction between zinc *coordination* and zinc *binding*. The first is specific coordination of zinc in the insulin hexamer, holding it together *via* the histidine residues. Zinc binding refers to the more aspecific binding of zinc, mainly to carboxylic acid groups on the surface of the hexamer. The most convincing site for zinc binding is in the centre of the hexamer between the B13 glutamic acid side chains. Binding of Cd^{2+}, Ca^{2+} and Pb^{2+} is tighter in that position than binding of zinc (Hill *et al.*, 1991). Simply soaking zinc ions into a 2Zn insulin crystal only increases aspecific zinc binding; the conformational change to the 4Zn form does not occur (Emdin *et al.*, 1980).

Because Zn^{2+} is spherical, the crystal field is undistorted and there is no real preference for either octahedral or tetrahedral coordination. In insulin, the conformation of the molecule (see chapter 4) dictates the coordination of the metal. In the native T6 hexamer, both zinc ions are octahedrally coordinated by three symmetry-related B10 histidine residues and three symmetry-related water molecules. These zinc ions are located in a shallow depression of the hexamer, and are easily diffused in and out of the hexamer. In the native T3R3 hexamer, the zinc coordination in the T3 trimer is octahedral, equivalent to that in the T6 hexamer. In the R3 trimer, however, both the axial and the off-axial zinc sites have tetrahedral coordination. The axial zinc ion is coordinated by the three symmetry-related B10 histidines of molecule 2, with the salt anion as fourth ligand. The off-axial zinc ion is coordinated by the second conformation of the B10 histidines of molecule 2, B5 His from a symmetry-related molecule 2 and two salt anions. With the different conformation of the N-termini of the B chains, the axial zinc ion in the R3 trimer is shielded from the environment by three α-helices, and the rate of diffusion of zinc from the hex-

amer is decreased, lowering the rate of dissociation of the hexamer itself. The 4Zn crystal form, when used in pharmaceutical preparations, therefore has slow-acting properties. In the native R6 hexamer, zinc ions are only found on the local threefold axis, tetrahedrally coordinated by three B10 histidines and usually a salt anion, if present in the crystallisation solution, as fourth ligand. Both zinc ions are covered by α-helices of residues B1 to B19 and therefore do not readily diffuse from the hexamer. As discussed in chapter 4, the R6 hexamer is therefore the most stable hexamer species in solution (Kaarsholm *et al.*, 1989).

6.2 The role of zinc in insulin biosynthesis

The zinc-free pig proinsulin hexamer has a net charge of -12 at neutral pH and has a capacity to bind 30 zinc ions (Grant *et al.*, 1972). Proinsulin is therefore positively charged and thus highly soluble, which is a necessary property as it has to travel through the endoplasmic reticulum and the Golgi. Most of the zinc binding occurs on the surface of the hexamer, where pairs of carboxylate groups pair up to bind zinc. The proinsulin C-peptide provides 24 of the carboxylates involved in this zinc binding. As the proinsulin is transported through the Golgi, conversion to insulin starts. In that process, the 24 carboxylate groups on the surface of the hexamer are lost, drastically reducing the capacity to bind zinc. At the same time, the 24 (6*4) connecting basic residues are removed. Thus the resulting insulin hexamer also has a net charge of -12 in the absence of zinc at neutral pH. When an insulin hexamer binds six zinc ions (the coordinated ions in the centre included), the resulting complex is neutral and precipitates. The insulin hexamer is very insoluble, and aggregates in microcrystalline form at pH 5.2-5.6 (Schlichtkrull, 1958). The insulin storage vesicles are therefore very densely packed with crystalline insulin, which reduces the susceptibility to enzymes that would break down the protein.

6.2.1 Charges

The effect of charges and the resulting net charge is obviously very important in the production of insulin in the body. Modifications of the insulin molecule involving a change in charges must therefore have some serious implications in insulin therapy. Most of the modifications introduced in the insulins described in this thesis affect the charge, as shown in table 6.1.

insulin name	modification	charge change	net charge of resulting monomer
B13Q	Glu→Gln	one fewer negative	-1
B8S/B13Q/B30amide	Gly→Ser	neutral	
	Glu→Gln	one fewer negative	
	carboxylate→amide	one fewer negative	0
B9D/B27E	Ser→Asp	one extra negative	
	Thr→Glu	one extra negative	-4
B25Y(B29-A1)A4Q	Phe→Tyr	neutral	
	delete B30 carboxylate	one fewer negative	
	Glu→Gln	one fewer negative	0
A21G/B9E/B10E	Asn→Gly	neutral	
	Ser→Glu	one extra negative	
	His→Glu	one extra negative	-4
B9H	Ser→His	neutral	-2
CoI	Zn^{2+}→Co^{2+}	neutral	-2

Table 6.1: Calculated net charge of the modified insulin monomers, at neutral pH

In crystallisations various compounds are added, such as precipitant, buffer, solvent and cryoprotectant. Any one of them might have an influence on the net charge of the polymer species in the crystal. Although it is difficult therefore to determine the exact charge of the crystal species, and despite the resolution of crystal structure determination often not being sufficient to determine the occupancy of ions, it is still possible to discuss the effects of specific mutations (see table 6.2). The B9D/B27E mutant is given here as if it has the pH of the buffer, namely pH 6. As the crystallisation was probably at lower pH, there would have been up to four fewer negative charges in the resulting dimer.

insulin name	crystal species	known additives†	net charge
zinc-free	hexamer	–	-12
2Zn	hexamer	2 Zn^{2+}	-8
4Zn	hexamer	2.67 Zn^{2+}, 1 Cl^{-}‡	-7.67
monoclinic	hexamer	2 Zn^{2+}, 2 Cl^{-}	-10
B13Q	hexamer	4 Zn^{2+}, 2 Cl^{-}	0
B8S/B13Q/B30amide	hexamer	2 Zn^{2+}, 2 Cl^{-}	+2
B9D/B27E	dimer	–	-8
B25Y(B29-A1)A4Q	hexamer	2 Zn^{2+}	+4
A21G/B9E/B10E	dimer	–	-8
B9H rhombohedral	hexamer	7 Zn^{2+}, 2 Cl^{-}	0
B9H monoclinic	hexamer	7 Zn^{2+}, 6 Cl^{-}	-4
CoI	hexamer	2 Co^{2+}, 1 I^{-}, 1 Na^{+}, 1 Cl^{-}	-9

Table 6.2: Calculated net charge of the aggregated insulin species, at neutral pH.
† modelled in electron density and (partly) refined;
‡ from Smith *et al.* (1984)

A net charge of zero for a hexamer with ligands would readily result in precipitation or crystallisation. Therefore, the net charges for B13Q insulin and B9H insulin in rhombohedral crystal form indicate a stable species. On the other hand, the strongly negative charge of the monomers of B9D/B27E and that of A21G/B9E/B10E combined with the removal of the main potential zinc binding site, leaves the lack of hexamerisation as no surprise. Both those mutants would bring 24 negative charges into a hexamer species, which would require binding of 12 metal ions. A hexameric species of B9D/B27E has been described by J. Turkenburg (1992). It was crystallised at neutral pH in the presence of 1M calcium and phenol. Upon careful inspection of electron density maps, six peaks in the core of the hexamer were found to have higher coordination than possible for water molecules, and were subsequently modelled as Ca^{2+} ions, reducing the net charge of the hexamer to -12, presumably with further disordered counterions.

The stability of the hexamer in the crystal obviously does not only depend on

the charges present. In monoclinic insulin, where the hexamer structure is R6, the stabilising effect of the zinc coordination itself is enhanced by the inability of the zinc ions to diffuse out of the hexamer. Zinc-free insulin can be kept in crystal form as a hexamer, even without the zinc coordination in the core to neutralise (some of) the charges. The B13Q mutant has also been reported to be stable in crystal form as zinc-free hexamers (Bentley *et al.*, 1992). Additionally, most crystallisation solutions contain a certain amount of metal ions other than zinc, such as calcium or sodium. Detection of these in crystal structures is difficult because they do not add enough electrons to the electron density maps and may not bind tightly or specifically enough. Sometimes the number of ligands coordinating to a peak gives an indication for metal binding. The aspecific binding of metal ions on the surface of the protein is difficult to detect because the coordination number associated with a specific electron density peak is often uncertain, as protein residues on the surface of the molecule are often disordered or so flexible that they can not be modelled.

6.3 The influence of phenol

Phenol was first detected in the monoclinic insulin hexamer (Derewenda *et al.*, 1989) in a position previously seen occupied by zinc in the 4Zn T3R3 structure (in the off-axial site), and B6 Leu in the 2Zn T6 structure (see figure 6.1). Phenol and phenol-like compounds were originally added to insulin preparations as anti-bacterial agents. As it turns out, phenol has a dramatic effect on the insulin conformation, driving it to the R-state. Depending on the concentration of phenol present, a hexamer may become T3R3 at low concentrations or R6 at high concentrations. Since phenol has a direct effect on the insulin conformation, it has an indirect effect on the coordination of the metal ions.

The phenol concentration in the monoclinic crystallisation protocol is 0.67%, resulting in R6 hexamers at pH 6.5 to 7.8. The B8S/B13Q/B30amide mutant crys-

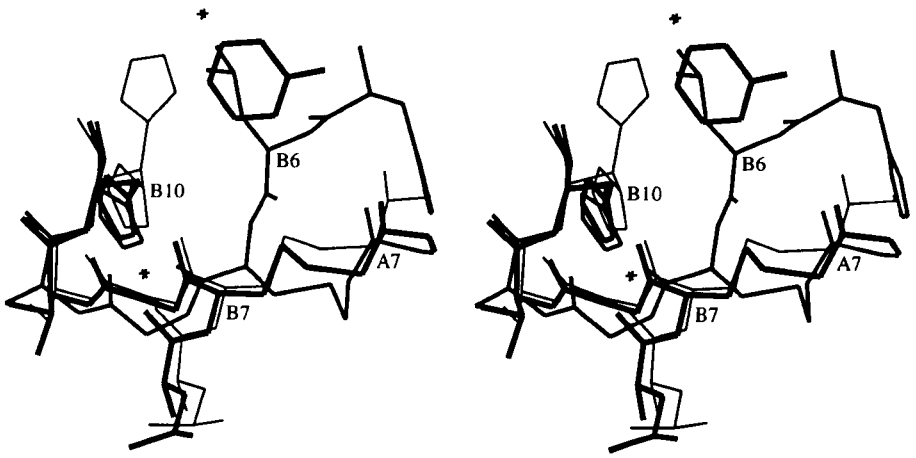

Figure 6.1: The phenol pocket region in three native insulin structures. Overlap of molecules 2 of T3R3 (thin lines) and T6 (medium lines) rhombohedral insulin onto R6 monoclinic insulin (thick lines) showing the region occupied by phenol in R6 insulin, zinc in T3R3 insulin and B6 Leu in T6 insulin.

tals were grown according to this protocol, resulting in monoclinic crystals and R6 hexamers, with two zinc ions on the local threefold axis, tetrahedrally coordinated by three B10 histidine residues and a chloride anion. B9H insulin was also crystallised following the monoclinic protocol, with essentially the same results. Because of the presence of additional histidines, more zinc was found in the core of the hexamer (see section 6.4).

Upon varying phenol concentration and pH, different crystal forms and different hexamers are produced. Whittingham *et al.* (1995) have reported a rhombohedral crystal form with $T3R3^f$ hexamers with a phenol concentration of 0.05% in the crystallisation solution, at pH 5.2-5.5. Smith and Dodson (1992) have described a rhombohedral R6 insulin hexamer that binds phenol, crystallised in the presence of 1% phenol at pH 8.49. CoI insulin was crystallised with 0.67% phenol, an equivalent amount to the monoclinic protocol. It did not, however, yield monoclinic crystals. The rhombohedral crystals produced contain $T3R3^f$ hexamers. It appears that the

preference of Co^{2+} ions for octahedral coordination has resisted phenol's drive to the R6 conformation, regardless of the presence of sufficient phenol. In addition to that, the phenol site contains two phenol molecules instead of one, as shown in figure 3.38, further stabilising the Co^{2+} forced in a tetrahedral conformation. Most complexes with phenol have only one ligand per monomer.

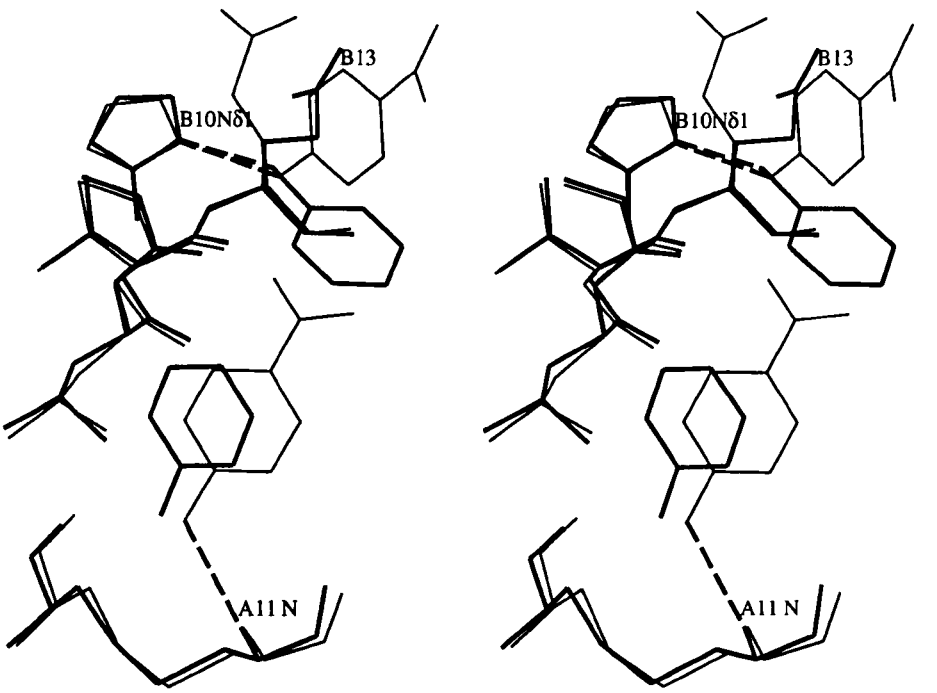

Figure 6.2: Stereo view of the double ligand site in CoI insulin and the insulin/4-hydroxybenzamide complex (Smith *et al.*, 1996), showing the phenolic derivatives

Smith *et al.* (1996), however, also describe a complex of insulin with two ligands per monomer, in which case the ligand is the phenolic derivative 4-hydroxybenzamide. One ligand is in the usual position, hydrogen bonding to the main chain of A6 and A11 of molecule 2. The other ligand is directly hydrogen bonded, through its hydroxy group, to Nδ1 of B10 His in molecule 2 and the carbonyl group of the first ligand. There are also indirect hydrogen bonds, through a water molecule, of the

amide/carbonyl function of the second ligand with the main chain carbonyl and Nδ1 of B10 His in molecule 1 in a symmetry-related dimer. In CoI insulin the hydrogen bonds of the first ligand are conserved, although that with the main chain nitrogen of A11 is slightly long (3.4Å). The second phenol molecule has a hydrogen bond to Nδ1 of B10 His in molecule 2, just like the second 4-hydroxybenzamide. This phenol molecule also has a short hydrogen bond to the side chain of B13 Glu in molecule 2. The position of the second ligand is quite different, as can be seen in figure 6.2. The second phenol molecule takes up a position partly filled by the amide group in the benzamide ligand, which means that the B13 Glu residues are able to take up different positions in the different structures. The bridging to another dimer does not occur in the double-phenol case, since there is no appropriate functionality on the *para*-site of the phenol.

6.4 Coordination of the metal

The T-conformation in the insulin hexamer is linked with a site of octahedral metal coordination on the (local) threefold axis, whereas the metal coordination in insulin in the R-conformation is always tetrahedral, whether on or off the axis, because of steric hindrance. In spectroscopic studies, the substitution of Co^{2+} for Zn^{2+} introduces a sensitive chromophoric probe of the structural and chemical events during the T $\rightarrow$ R transition. Ciszak and Smith (Ciszak and Smith, 1994) argue that a certain amount of care should be taken over this, since there is evidence of both octahedral and tetrahedral coordination around the zinc in the T trimer of the T3R3^f structure they describe. This would lead to misinterpretation of the extent of the T $\rightarrow$ R transition based on spectroscopic evidence alone. Following the discussion of the preference of Co^{2+} for octahedral over tetrahedral coordination, some consideration needs to be given to the applicability of the ion in the study of the phenomenon of T $\rightarrow$ R transition in insulin. The conditions under which Co^{2+}

insulin converts from T to R conformation seem to have to be more extreme than is necessary for Zn^{2+} insulin.

The CoI insulin described in this thesis is in T3R3^f conformation and has the expected octahedral metal site with coordination by three symmetry-related B10 histidines from molecule 1 and three water molecules. The other metal site is tetrahedral, with the three symmetry-related B10 histidines from molecule 2 and an iodide ion as ligands. The histidines in the R3^f trimer are stabilised by an extra phenol ligand in the phenol pocket. The presence of the iodide further stabilises the essentially unfavourable coordination of the Co^{2+} ion.

The B13Q mutant, also in T3R3^f conformation, has quite a different distribution of metal in the hexamer. The metal site in the T3 trimer is as expected, octahedrally coordinated by three symmetry-related B10 His residues from molecule 1 and three water molecules. In the R3^f trimer, the B10 histidines of molecule 2 have fully changed their side chain conformation to the second alternative seen in native T3R3^f insulin. Hence the mutant has three fully occupied tetrahedrally coordinated zinc ions in the off-axial sites, as shown in figure 6.3. Bentley *et al.* (1992) have postulated that the Glu to Gln mutation in itself favours R-conformation, without the need for zinc ions or halides. The requirement of the mutant for Zn^{2+} in the core of the hexamer to balance the negative charges is absent, which might explain the total change in conformation of the B10 histidine side chain in the R3^f trimer away from the position where it was needed in the native structure to coordinate zinc.

In B8S/B13Q/B30amide insulin, in R6 conformation, both zinc sites on the local threefold axis have tetrahedral coordination with three B10 histidines, and probably chloride as a fourth ligand. The T3R3 form of the mutant has been described by Dodson *et al.* (1991). Although it has the same mutation as B13Q insulin, and thus the same non-requirement for zinc in the core of the hexamer, there is no indication of off-axial zinc binding in the B8S/B13Q/B30amide mutant. This might be related to the different conformations of the side chains of the B13 residues (see

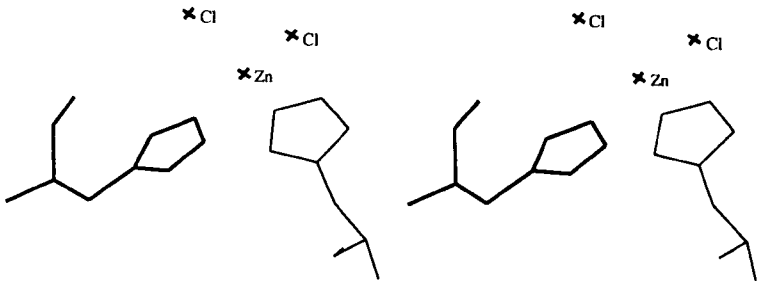

Figure 6.3: Stereo view of off-axial, tetrahedrally coordinated zinc in B13Q insulin. B10 His in thick lines, symmetry-related B5 His in thinner lines.

chapter 4).

In the B9H mutant, in both forms, rhombohedral T3R3^f and monoclinic R6, there is additional zinc present in other sites than the axial sites coordinated by B10 histidines. In the mutant T3R3^f hexamer, the zinc ions are distinctly organised. In the T3 trimer the axial zinc ion is unexpectedly tetrahedrally coordinated by three symmetry-related B10 His residues from molecule 1 and chloride. The electron density for the fourth ligand is not threefold symmetrical and has therefore been modelled and refined as a chloride. On the other side of the B10 His side chain, a peak was found in which an additional zinc ion could be modelled, and refined with an occupancy of 0.67. This zinc ion is tetrahedrally coordinated by the Nδ1 of the B10 His, the Nδ1 of a symmetry-related B9 His, and two water molecules (see figure 3.32). The geometry of the B9 His residue with respect to the zinc ion is imperfect, although the coordination of the zinc ion is reasonable at first glance. The tetrahedral angles around the zinc ion range from 97.3 to 123.2°. In the T3 trimer of the native T3R3 hexamer, this region is essentially taken up by water structure, with the side chain of residue B13 from molecule 2 folded slightly more towards the T3 trimer. In the mutant the carboxylate group of this residue is folded away, towards its own trimer. In the R3^f trimer, the axial zinc ion has regular tetrahedral

coordination with three symmetry-related B10 histidines from molecule 2 and a chloride anion, equivalent to the native structure. The other zinc coordination site is in almost the same position as the off-axial zinc in the native structure (the sites are less than 1Å apart). However, instead of the second conformation of the B10 His of molecule 2 in the native structure, the histidine residue at position B9 is involved in the tetrahedral coordination of the zinc ion, together with B5 His of the same monomer. Both histidines have Nϵ2 directed towards the zinc ion. The other two ligands are water molecules. The geometry of the histidine side chains with respect to the zinc ion and the tetrahedral coordination are almost perfect (see tables 3.17 and 3.18), and the main chain conformation is totally conserved as compared to the native structure. The side chain conformation of B13 Glu in molecule 1 is also equivalent to that seen in the native T3R3 structure.

In the monoclinic R6 form of the B9H mutant, the zinc sites on the local threefold axis are as expected: tetrahedral coordination with three B10 histidines and a chloride anion. In an overlap of all six B chain α-helices of the mutant on the corresponding helices of the native structure, one trimer has significantly better overlap than the other. The zinc ions connected with the second trimer in the mutant are almost 1Å away from the zinc positions in that trimer in the native structure. This means that building up the mutant hexamer from dimers positioned separately by molecular replacement was the best approach. There are four zinc coordination sites involving the Nδ1 atoms of B10 and B9 histidines, all modelled with chloride anions as third and fourth ligands, and refined with half occupancies for the zinc and chloride ions. The coordination of these zinc ions is essentially tetrahedral, although the geometry of the histidines with respect to the zinc ions is very unusual. The term 'coordination' is perhaps best not used in such cases. In addition to those zinc sites, there is a pool of density in the centre of the hexamer. With the aid of a peak searching program (X-SOLVATE (Molecular Simulations, 1996)), six of the peaks could be modelled as zinc ions, with Nϵ2 atoms of B9 His and Oϵ atoms of B13 Glu

residues in the vicinity (less than 2.5Å). The coordination of these six zinc ions is irregular, indicating disorder, and they were refined with half occupancy.

The B25Y(B29-A1)A4Q mutant has not been fully refined, and the resolution of the data does not allow detailed analysis of the coordination of the (putative) zinc sites. There is, however, no reason to doubt octahedral coordination in this particular T6 hexamer.

6.5 Other metals in other places

There is evidence that Ca^{2+} ions are taken up by rhombohedral insulin crystals *in vivo* (Howell *et al.*, 1978). Hill *et al.* (1991) have described an attempt to detect Ca^{2+} ions in a 2Zn insulin crystal, which failed because the calcium ions are not easily distinguishable from water molecules. In the monoclinic form of the B9D/B27E mutant, described by J. Turkenburg (1992), calcium ions were found to balance the charges in the core of the hexamer. In that case the number of ligands involved was too high for a water molecule.

For the modified insulins described in this thesis, alternative metal binding is only proposed for the CoI insulin. One region of electron density in the $R3^f$ trimer on the threefold axis on the surface of the hexamer, in the channel left by the extended conformations of the three N-terminal residues of the B chains of molecule 2, could not be modelled satisfactorily with water molecules. Attempts were made to model a threefold disordered citrate ion or acetate ion on the threefold axis, bonded to a sodium ion off the threefold axis. The sodium ion refined reasonably well, and had a water molecule 2.0Å away (see figure 6.4). The modelling of an acetate ion was promising, although the refinement was unsuccessful, and the ion was not included in the calculations of the final maps and R-factors.

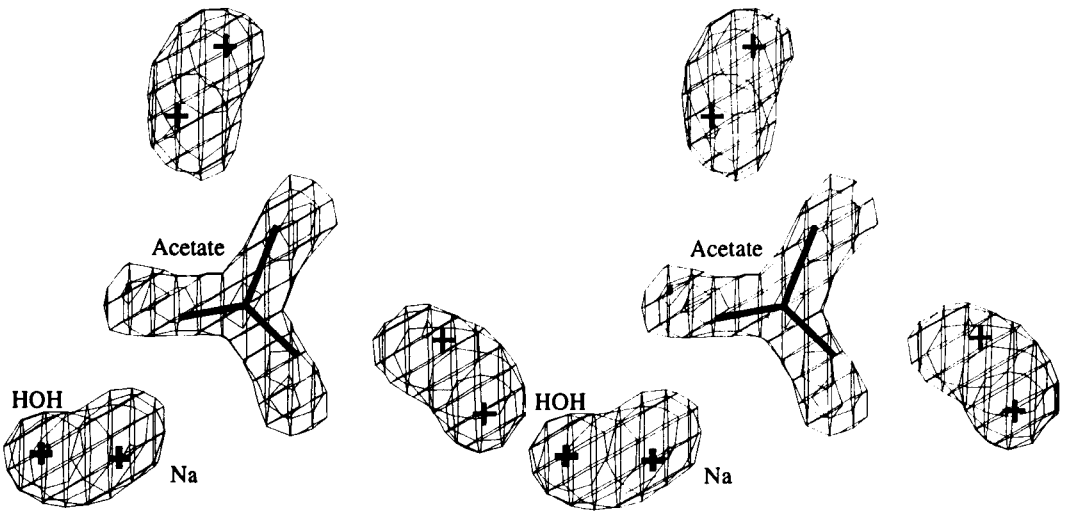

Figure 6.4: Sodium acetate in CoI insulin. Stereo view down the threefold axis. $2F_o-F_c$ electron density at $+1\sigma$ for sodium and water positions, F_o-F_c electron density at $+3\sigma$ for the acetate position.

6.6 Summary

Metal binding and metal coordination in modified insulins have been studied on the basis of existing knowledge. Binding of zinc is generally accepted to be the driving force of hexamerisation, although metal-free hexamers can be formed in certain circumstances. The occurrence of octahedral and tetrahedral coordination of the metal is linked to the T-state and R-state trimers, respectively, although one exception to this rule was seen in one of the mutants. The presence of phenol is also involved in this connection between coordination and conformation. The detection and refinement of other metal sites in crystal structures is difficult for various reasons.

Chapter 7

Conclusions and future work

7.1 Conclusions

The first and foremost conclusion of structural studies of insulin must be that the protein is extremely flexible and adaptable. Even the native protein can quite easily be crystallised in numerous crystal forms depending on their specific T and R conformations. The flexibility is especially marked in the B chain: the conformation of the N-terminus gives rise to the T and R naming system, and the flexibility of the C-terminus is thought to be very important in a conformational change necessary in receptor binding.

Because the protein is relatively small, even a small change in one of the chains constitutes a relatively large difference in the structure. For example, one point mutation represents almost 2% of the protein. This might explain why two isomorphous structures of insulin may still have an isomorphous R-factor significantly higher than 10%. As another example, the flexible termini of the B chain (B1-B3 and B29-B30) represent around 10% of the scattering matter (excluding solvent) in the X-ray diffraction experiment, and thus their flexibility has a negative effect on the quality of diffraction.

At the same time, insulin's flexibility provides interesting and unexpected be-

haviour. Mutants designed to be monomeric in pharmaceutical preparations may still crystallise as dimers or even higher aggregates. Unexpected metal binding is seen, both in the position and the coordination of the metal ion. The conformation of the B chain is not always predictable.

The protein's flexibility and adaptability is and will be used in the formulation of therapeutic preparations. The adaptability of insulin to its environment can be used to produce highly stable forms which can be stored for a length of time (*e.g.* by incorporating certain additives), after which the protein may change to a form with a specifically desired activity upon administration to a patient.

A few specific topics have been raised through the studies of the modified insulins described, namely:

- The T3R3 hexamer, with the B chain N-terminus of one of the molecules in the asymmetric dimer fully helical, probably does not exist. Even the native 4Zn insulin structure is most likely $T3R3^f$, where residues B1-B3 in molecule 2 of the dimer are extended. This structure will be re-refined in the near future;

- The use of ethanediol in insulin crystallisation instead of the very volatile solvents ethanol and acetone seems very promising, and should be encouraged in cases where cryocrystallographic experiments are anticipated;

- The eight structures described in this thesis have all been refined in different ways. Even in cases where it was tried to use a similar protocol for similar structures, success was limited. A general protocol for insulin refinement would have to take into account especially the observation:parameter ratio. Insulin refinement often involves a limited number of reflections, which causes problems with the reliability of the statistics, especially those of the free R-factor.

7.2 Future work

Following the investigations described in this thesis, a number of projects could be taken up, involving the modified insulins studied:

- Further refinement of rhombohedral B9H insulin when REFMAC has been adapted to deal effectively with special positions and special distances involving symmetry-related atoms;

- Further data collection and refinement of B9D/B27E insulin in orthorhombic crystal form;

- Renewed attempts to obtain better crystals of B25Y(B29-A1)A4Q and A21G/B9E/B10E in order to obtain higher resolution data and better structural information;

- It would be interesting to obtain other crystal forms (to study other conformations) of some of the modified insulins, especially:

    1. CoI in all-R-conformation (by using more phenol or a more powerful ligand in the crystallisation), to see if the cobalt ions can both be in tetrahedral coordination;

    2. B8S/B13Q/B30amide in all-T-conformation, to investigate further the unusual T-conformation seen in the low salt form;

    3. B9H in all-T-conformation, to see if the mutant could accommodate 6 zinc ions in that conformation;

- Further study of the conformational preference of the B13Q mutant and its unusual behaviour in the presence of thiocyanate, a ligand not previously seen to allow all-R-conformation. An attempt has been made to crystallise B13Q insulin with thiocyanate, resulting in a rhombohedral T3R3 crystal form (J.

Whittingham, pers. comm.). A higher concentration of thiocyanate might produce the desired monoclinic R6 form;

- It would be of interest to attempt to crystallise some of the modified insulins in the absence of zinc (or closely related ions like cobalt) but in the presence of calcium. This might be especially revealing for A21G/B9E/B10E in which case calcium could balance the charges at the dimer interface, while the zinc-binding functionality is removed at B10;

- The insight into the flexibility obtained from these studies and this flexibility itself should be used in binding studies of the protein with its receptor, and in crystallographic studies of the receptor complex.

Appendix A

Lagrange multipliers

A.1 A general introduction to Lagrange multipliers

The idea for the Lagrange multipliers was first put forward in the ninth of Lagrange's "Leçons sur le calcul des fonctions" in 1806. A function $M = f_0(x_i)$ for $i = 1, 2, ...N$ is to be maximised, subject to constraints g_j for which is known:

$$g_j(x_i) = 0 \qquad \text{for } j = 1, K \tag{A.1}$$

This is equivalent to reducing the dimension of the original system.

A new function L is constructed, adding the dimension lost through the constraints by incorporating Lagrange multipliers. Therefore, the constrained maximisation of function M is equivalent to the unconstrained maximisation of function L where

$$L = f_0(x_i) + \sum_j \lambda_j g_j(x_i) = f_0(x_i) + \lambda_1 g_1(x_i) + \lambda_2 g_2(x_i) + ... \tag{A.2}$$

This function has $N + K$ unknowns.

The maximisation is performed by differentiating L with respect to both parameters. In the maximum the differentials with respect to the appropriate λ_k and

x_k will be 0:

$$\frac{\partial L}{\partial \lambda_k} = g_k(x_i) = 0 \tag{A.3}$$

and

$$\frac{\partial L}{\partial x_k} = \frac{\partial f_0}{\partial x_k} + \sum_j \lambda_j \frac{\partial g_j}{\partial x_k} = 0 \tag{A.4}$$

We now have $N + K$ equations (equations A.4 and A.3, respectively) to solve the $N + K$ unknowns of function L.

A.2 Lagrange multipliers in entropy maximisation

In the general case of maximising entropy, the following entropy formula may be used (this is equation 2.57):

$$S(\mathbf{p}, \mathbf{q}) = -\sum_{i=1}^{B} p_i ln(p_i/q_i) \tag{A.5}$$

The constraints may be denoted as follows:

$$\sum_{i=1}^{B} p_i C_{ij} = c_j \tag{A.6}$$

for B possible system states. $j = 1, M$ where M is the number of constraints. Since the constraints must always be satisfied, it follows from equation A.6 that

$$\sum_{i=1}^{B} \Delta p_i C_{ij} = 0 \tag{A.7}$$

where Δp_i is a small variation in p_i. In the maximum of S we have:

$$\sum_{i=1}^{B} \frac{\partial S}{\partial p_i} \Delta p_i = 0 \tag{A.8}$$

Now a multiplier λ_j is introduced for each constraint. Multiplying equation A.6 by λ_j for each j and summation of the results leads to

$$\sum_{i=1}^{B}\sum_{j=1}^{M}\lambda_j \Delta p_i C_{ij} = 0 \tag{A.9}$$

When $M \leq B$, only $B-M$ parameters Δp_i are independent, whereas the parameters for which $i = 1, M$ are dependent. This means the multipliers λ_j may be obtained from

$$\frac{\partial S}{\partial p_i} = \sum_{j=1}^{M}\lambda_j C_{ij} \tag{A.10}$$

A.3 Lagrange multipliers in crystallography

In crystallography the entropy may be maximised in the case of general linear constraints of the form

$$\int_V p(\vec{r})C_j(\vec{r})d\vec{r} = c_j \tag{A.11}$$

for $j = 1, M$ where M is the number of constraints. $C_j = e^{2\pi i \vec{h}_j \cdot \vec{r}}$ for every reciprocal lattice vector $\vec{h}_j$, and $c_j = U(\vec{h}_j)$ is the unitary structure factor for the reflection corresponding to $\vec{h}_j$. $p(\vec{r})$ is the unknown optimal distribution of atoms and V the volume of the unit cell. The normalisation condition, or the origin term with $j = 0$, may also be regarded as a constraint:

$$\int_V p(\vec{r})d\vec{r} = \int_V p(\vec{r})C_0(\vec{r})d\vec{r} = c_0 = 1 \tag{A.12}$$

The constrained maximisation of the entropy $S(\mathbf{p}, \mathbf{q})$ (see equation A.5) is equivalent to the unconstrained maximisation of

$$S(\mathbf{p}, \mathbf{q}) + \sum_{j=0}^{M}(\lambda_j \int_V p(\vec{r})C_j(\vec{r})d\vec{r} - c_j) \tag{A.13}$$

where λ_j are the Lagrange multipliers. $\mathbf{q} = q(\vec{r})$ is the prior distribution.

Differentiation with respect to λ_j returns the constraint equations A.11, while

differentiation with respect to $p(\vec{r})$ yields:

$$-1 - ln(p^{ME}(\vec{r})/q(\vec{r})) + \sum_{j=0}^{M} \lambda_j C_j(\vec{r}) = 0 \tag{A.14}$$

or

$$p^{ME}(\vec{r}) = q(\vec{r}) exp(\lambda_0 - 1) exp(\sum_{j=0}^{M} \lambda_j C_j(\vec{r})) \tag{A.15}$$

where p^{ME} is the Maximum Entropy posterior distribution. Solving directly for λ_0 by integrating this equation over $\vec{r}$ and using the normalisation constraint $\int_V p^{ME}(\vec{r})d\vec{r} = 1$ gives $\lambda_0 - 1 = -lnZ$ where

$$Z(\lambda_1, ..., \lambda_B) = \int_V q(\vec{r}) exp(\sum_{j=1}^{M} \lambda_j C_j(\vec{r}))d\vec{r} \tag{A.16}$$

Z is the partition function which describes the way in which the unit cell has been divided up in fractions B. Then

$$p^{ME}(\vec{r}) = \frac{q(\vec{r})}{Z(\lambda_1, ..., \lambda_B)} exp(\sum_{j=1}^{M} \lambda_j C_j(\vec{r})) \tag{A.17}$$

and

$$\int_V p^{ME}(\vec{r}) C_j(\vec{r}) d\vec{r} = c_j \tag{A.18}$$

Appendix B

Laboratory safety and reproduction

B.1 Introduction

Work in a laboratory involved in protein crystallographic research can pose extra risks for pregnant women compared to the average individual. The risks are multiple and varied, and some are important even before conception. Moreover, documentation is available on the risks for foetuses brought on by the male parent, both before and after conception. For example, it has been reported that exposure of males to anaesthetic gases can lead to adverse pregnancy outcome (mainly increased incidence of congenital anomalies in the offspring) in their unexposed wives (Strobino *et al.*, 1978). It has also been suggested that the male can contribute to the incidence of malformations in foetuses if substances absorbed in his body are secreted into the seminal fluid, subsequently absorbed in the vaginal mucosa during intercourse, after which they can affect a developing embryo or foetus (Clement Associates, 1981).

In the laboratory in York the research can be divided into three spatially separated sections according to the type of laboratory involved:

1. Molecular biology/ biochemistry

(a) cloning of genes coding for the desired protein

(b) expression, isolation and purification of the protein

(c) crystallisation

(d) crystallisation and/or soaking of heavy atom derivatives

2. X-ray
   data collection

3. Computing
   data processing, structure solution and refinement

Each of the above sections and subsections is connected with specific risks to reproduction, whether the growing foetus or the parents, either before or after conception.

B.2 Teratogenicity

Teratology is the study of malformations or serious deviations from the normal type in organisms (Webster's, 1965). A teratogenic agent is defined by its ability to produce a congenital defect (whether physical or functional) in the conceptus or offspring (Shepard, 1980).

Some biological and chemical hazards are easily linked with teratogenicity. The effects of mutagens, carcinogens and radiation (whether X-rays, microwaves or radioisotopes) on either parent are less obvious candidates.

B.3 Biological hazards

During any period of suppressed immunological response, and thus during pregnancy, a person is more susceptible to infections. This is especially true for viral infections, which have been proved to have disastrous consequences for human foetuses. In an average molecular biological or biochemical laboratory, like in York, no

pathogenic strains of viruses, bacteria or fungi are used, so primary (direct) or secondary (indirect: the male infecting his wife after being infected at work) infections are unlikely. The risk of contracting viral diseases, like rubella (german measles) or varicella (chicken pox), elsewhere, is far greater.

In addition to being infectious, aetiological agents as described above, may produce illness after ingestion of certain cell wall components (endotoxins) or exotoxins, or may cause an allergic response after handling them.

A second type of biological hazard involves research with oncogenic viruses, which are capable of inducing tumours or cancer after infection. In most of the literature consulted, it is suggested that pregnant women or women who are planning to get pregnant, should not be allowed to work with oncogenic viruses (see for example Rayburn (1990) p.129 and p.298).

More relevant to the situation in York, are the hazards involved in recombinant DNA research. In the early days of the use of the technique a moratorium was proposed (Berg *et al.*, 1974) as a result of serious concerns that artificial recombinant DNA molecules could, after introduction in as common a bacterium as *E. coli*, "become widely disseminated among human, bacterial, plant or animal population with unpredictable effects". It was suggested that, until hazards were better understood and appropriate precautions could be taken to prevent the spread of "new" DNA, no experiments should be undertaken that involved

- introduction of antibiotic resistance or toxin formation in strains that did not carry such already, or construction of plasmids with combinations of antibiotic resistance not already existing in nature; or

- introduction of oncogenic viral DNA in autonomously replicating systems that could spread easily in humans and other species.

The consensus reached at a conference in Asilomar in 1975 formed the basis of the NIH (National Institutes of Health) guidelines, which are now widely accepted, though not enforced by any law. In support of recombinant DNA research one could state that the mechanisms for gene transfer, rearrangement and cloning already exist

in nature, that hazardous constructs would exist if they were possible, and that the chances of inadvertently creating a problem are small.

Retrospective studies have shown that the containment techniques proposed in the guidelines have been effective, with the absence of infections so far. However, containment may be against the very goal of certain experiments, where competition of "new" organisms with others in the environment is necessary.

As with the research involving oncogenic viruses, the better-safe-than-sorry option seems most appropriate for (potentially) pregnant women.

B.4 Chemical hazards

The teratogenicity of certain chemicals is well-documented (Shepard, 1980, Hunt, 1979, Rayburn, 1990). The mode of action of some of them, however, can be surprising, and is often ill-understood.

B.4.1 The male parent

Hunt (1979) devotes a whole chapter to the influence of environmental agents on male reproductive failure. The effects described include male infertility as well as adverse pregnancy outcome *via* mutation of the sperm. Two exposure routes may account for the effects:

1. exposure of the male results in exposure of the female, and the agent operates directly through the mother; or

2. exposure damages sperm cells which causes infertility or abnormal development of the embryo (the latter usually results in spontaneous abortion).

A number of other agents, like thalidomide, methadone, and certain therapeutic agents (epileptic drugs, antibiotics and anti-cancer drugs) are proved to be excreted into seminal fluid. There they may directly influence the sperm (see table B.1) or be absorbed in the vaginal mucosa after intercourse.

AGENT	EFFECT
afflatoxin B1†	reduced fertility
alcohol(chronic alcoholism)	testicular pathology
anaesthetic gases	increased incidence of congenital anomalies in offspring
anti-cancer drugs	reduced fertility
cadmium chloride†	reduced fertility
cigarette smoke	increase in abnormally-shaped sperm
glutamic acid†	reduced fertility
hydrocarbons	twofold increased incidence of childhood cancer
iodine(thyroid therapy)	atrophy of testes
lead	decreased sperm count and motility; increase in abnormally-shaped sperm
methylmercurychloride, methylmercury-hydroxide, mercury chloride†	reduced fertility
nickel sulphate†	reduced fertility
nitrous oxide†	suppressed spermatogenesis
urologic antibiotics†	altered spermatogenesis

Table B.1: Agents affecting sperm (adapted from Hunt (1979)).
†animal study only

B.4.2 The female parent

Similar to the discussion for the male parent, exposure to certain chemicals may cause infertility or gene mutations in the female. In addition, most teratogenic chemicals affect the developing foetus in the first trimester of pregnancy, often before the woman even suspects she has conceived. This leads some organisations to exclude all fertile women from work with such chemicals (Rayburn, 1990).

A developing embryo or foetus may be affected by chemicals through two routes:

1. some substances may produce toxic effects to the mother, thus interfering with placental function; or

2. other substances may cross the placenta, affecting the embryo/ foetus directly.

Adults have a reasonably effective detoxification system for ingested toxins. The chemical is diluted in the stomach and is often converted to lesser toxic substances. It is often selectively not absorbed by the intestine but first carried to the

liver to be detoxified. Unfortunately, the metabolites are sometimes even more toxic than the original substance. In the foetus and up to shortly after birth, the enzyme systems for detoxification may not have fully developed. This means that some toxic material is more dangerous, although the lack of development of the detoxification system may be an advantage in the case of more toxic metabolites.

Chemicals with a teratogenic effect are very diverse, and the teratogenicity of a newly synthesised chemical is not easily predictable. Moreover, teratogenic testing is more difficult than ordinary toxicity testing, and most information documented stems from accidental human exposure, anecdotal information and extrapolation from animal data (Safety Handbook, 1992).

For some substances the evidence is inconclusive (*e.g.* sulfonamides and cadmium) or the effects may be brought on or aggravated by other factors (*e.g.* coffee). Others seem to have no or only mild effects. Examples are: ammonium chloride, ascorbic acid, beryllium, chloramphenicol, chromium, cocaine, codeine, copper chloride, guanine and guanosine, heroin and penicillin. Obviously, the effect depends largely on the dose and the route of exposure. Many of the hazards mentioned in table B.2 do not represent an occupational exposure, but rather a problem in accidents. Most teratogens exert their effects in the early stages of pregnancy, during organogenesis, which starts with budding of the lungs, 28 days after conception. Formation of the heart membranes is completed after 46-47 days (Shepard, 1980), by which time the limbs have formed and bone formation has started. This means that the foetus is almost fully formed before the second missed menstrual period, a moment when many women would only begin to suspect to be pregnant.

AGENT	EFFECT
anaesthetic gases‡	spontaneous abortion, decreased birth weight, stillbirth, increased cardiovascular birth defects
cigarette smoke‡	spontaneous abortion, foetal growth retardation, placenta problems, increase in neonatal deaths, SIDS, lag in cognitive development
ethanol‡	reduced birth weight, increased still birth rates, increased congenital malformation rates
glycerol‡	lethal to mother upon injection
iodine/iodide	severe thyroid problems
lead‡	spontaneous abortion, decreased birth weight, prematurity, stillbirth, mental retardation
lithium chloride	cardiovascular malformations
methylmercury‡	gross injuries to central nervous system
nickel (salts or air pollution)	small defects, microsomia
phenylalanine	spontaneous abortion, heart disease
retinoids (vitamin A and similar compounds)‡	malformations of ear, palate, cardiac/aortic tissue, brain
salicylates	congenital defects, stillbirth
certain solvents(*e.g.* toluene‡	spontaneous abortion, certain rare congenital malformations, microsomia
spray adhesives	possibly mild effects upon inhalation
(dihydro)streptomycin	deafness
tetracyclin	staining of (deciduous) teeth
actinomycins†	various defects
adenine†	cleft palate, digital defects, haematomas
afflatoxin†	liver problems
aminoguanidine†	liver problems
monosodium L-aspartate†	various defects
sodium azide†	all tissues affected above certain concentration
5-bromo deoxyuridine†	polydactylism
cobaltous chloride†	anaemia
concanavalin A†	cranofacial, trunk and limb anomalies
2-deoxy adenosine/guanosine†	skeletal and palate defects
DMSO†	bone abnormalities
EDTA†	various defects
E.coli endotoxin†	hydrocephalus, neuronal necrosis
leucine†	various defects
pyridine†	muscular hypoplasia, micromelia
selenium†	microsomia, cranofacial defects
spermine†	total arrest of development
surfactants(*e.g.* SDS)†	effects at maternal toxic dose
tritium†	microsomia, microcephaly, small testes/ovaries
tryptophan†	various defects
uracil†	various defects
zinc chloride†	bone malformation

Table B.2: Female reproductive toxins. Adapted from Shepard (1980).
†animal study only, typically upon injection.
‡confirmed by animal study

Thalidomide is an example of a chemical that could be harmless and effective if used correctly. However, it took scientists just too long to discover that it is only the (+)-form that is safe, whereas the (-)-form is a potent mutagen. Unfortunately, the drug contained both forms, resulting in serious birth defects (Roberts, 1989).

In the Netherlands it has been decided (Loodbesluit; Ministerie van Sociale Zaken(1994)) that a pregnant woman can not be forced to work with lead and lead derivatives that can be taken up in the human organism, and that employers must provide alternative work.

B.5 Physical hazards

This section will include various non-biological and non-chemical influences on biological systems, in this case the human during reproductive years. The effects of radiation, temperature and work with Visual Display Units (VDUs) will be discussed.

B.5.1 Radiation

Two types of ionising radiation may be encountered in the laboratory in York: radioisotopes emitting β-particles, and X-rays. The radiation hazards of α-particles, γ-rays and neutrons will therefore not be discussed in detail. Figure B.1 shows the stages of radiation action in time. These times are dependent upon the system irradiated, and its temperature.

α-particles are helium nuclei produced by large unstable elements. Compared to other types of radiation, they are high in mass, charge and energy, and have a high specific ionisation, which means they are relatively efficient in creating new ion pairs on their way through matter. Although these properties make α-radiation so destructive when ingested, since all of the energy will be deposited in a small volume of tissue, on the other hand the particles have a very short range in air and are easily stopped by a shield or even the skin.

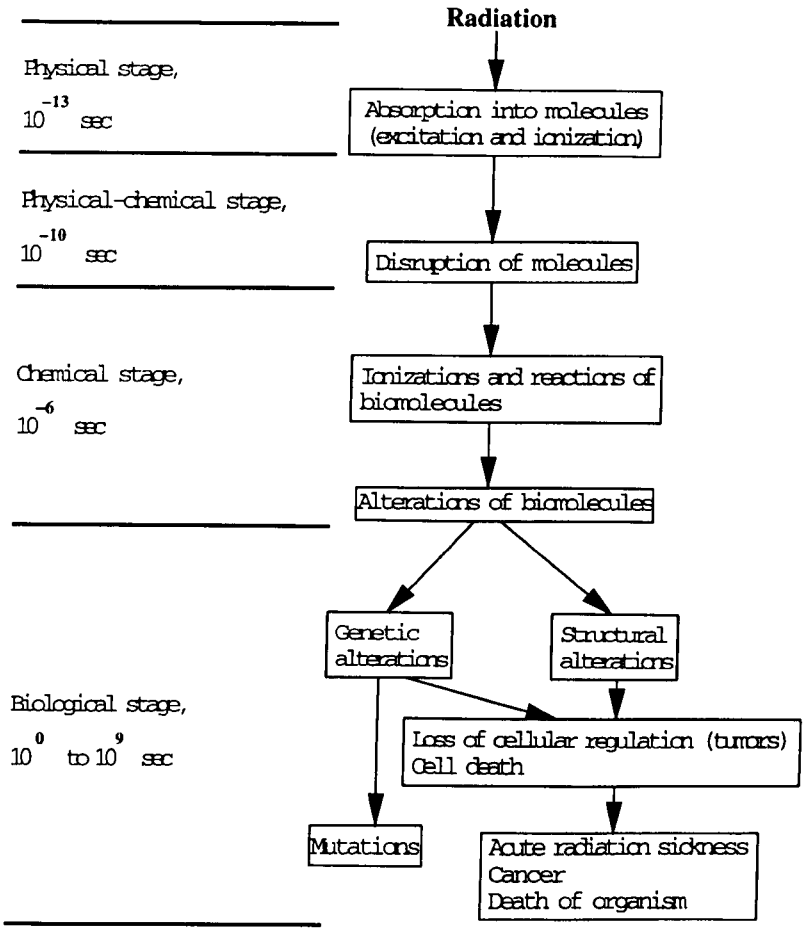

Figure B.1: Stages of radiation action with time. Reproduced from Rayburn (1990).

β-particles are unpaired electrons emitted by naturally occurring (^{3}H, ^{14}C, ^{40}K) or artificially produced (^{32}P, ^{35}S, ^{45}Ca) isotopes. Their range in air and other matter is directly proportional to their energy and the density of the substance (Local Rules, 1986). Although they are lower in energy than α-particles, they are also considerably smaller, which makes them equally hazardous inside the body, while the external hazard for high-energy β-emitters is even somewhat greater than that for α-particles. High-energy β-emitters have the unique characteristic of producing X-rays, specifically called 'bremsstrahlung', upon interaction with matter.

Atoms of higher atomic weight are more efficient bremsstrahlung-producers than those of lower atomic weight, so that lead, its usefulness as X-ray shielding recognised as early as 1903 (Fuscaldo *et al.*, 1980), is not the obvious material for β-shields, but plastic. Two other options for radiation hazard control are at the disposal of people working with β-emitters: keeping a distance and reducing exposure time. The intensity of the radiation drops rapidly, following the inverse square law, as one moves away from the source. This means that for low-energy β-emitters like ^{14}C and ^{35}S the dose rate is near zero at a separation of as little as 10cm. Since the total dose received is simply (dose rate * time), reducing exposure time is a very effective measure in decreasing exposure to high-energy β-emitters like ^{32}P.

γ-radiation and X-rays are types of electromagnetic radiation with identical physical characteristics but different manner of production. γ-rays arise from the excited nucleus of certain radioisotopes or reactor-fission-products, as it drops to a lower energy state, and often accompany the emission of α- or β-particles. X-rays are produced by slowing down of fast electrons in the electron cloud of high atomic-number atoms, or during the acceleration of charged particles. Because of the lack of mass and charge, γ- and X-rays have few interactions per unit volume of matter, which makes them far more capable of penetrating matter than α- or β-particles. They do not undergo electrostatic interactions. Instead, they interact with atoms to produce electrons which in turn can produce secondary ionisations, which are extremely penetrating. Therefore, γ- and X-rays pose a significant external as well as internal hazard. The three factors for limiting external exposure, namely time, distance and shielding (in this case lead), are all reasonably effective in the case of γ- and X-rays. The dose intensity decreases with the inverse square of the distance from the source. In addition, the penetration of γ- and X-radiation declines exponentially in matter, with the efficiency of attenuation dependent upon density and electron density of the absorbing material, lead, steel and concrete being used most often.

Neutrons are produced with accelerators, atomic reactors, spontaneously fissioning isotopes and by interaction of α- or γ-radiation on beryllium or certain other elements. Exposure to neutrons would cause considerable damage to biological sys-

tems, but fortunately neutrons are very rarely used in biomedical research laboratories. Research with neutrons would require specialist training and the availability of a professional safety expert.

In the descriptions of ionising radiation above, terms like dose rate and dose intensity have been used. Definitions of these and other radiation-related terms can be found in the University's Local Rules (1986). The Local Rules also give guidelines on the amount of radiation acceptable as occupational exposure, see table B.3.

Exposed part of body	Dose limit	
	calendar year	calendar quarter
whole body	50 mSv (5 rem)	
individual organs (skin, hands, forearms, feet, ankles)	500 mSv (50 rem)	
lens of the eye	150 mSv(15 rem)	
abdomen of women of reproductive capacity		13 mSv(1.3 rem)
abdomen/foetus of pregnant woman	10 mSv (1 rem) between declaration of pregnancy and delivery	

Table B.3: Occupational exposure limits according to University of York Local Rules (1986)

For members of the general public the limits are 1/10th of the limits in table B.3. Three other sources, of which only two are more recent, do not agree with the amounts in table B.3. Rayburn (1990) treats the skin separate from the other individual organs, with 30 rem and 75 rem as their maximum annual permitted doses, respectively. Also, a whole body dose of up to 3 rem per quarter is permitted, provided the annual dose would not exceed 5 rem. More worryingly, he states a total of 0.5 rem for the abdomen/foetus of a pregnant woman during the entire gestation period. Following on from this, fertile women should only be employed in positions where the annual dose is not likely to exceed 2 to 3 rem to prevent excessive exposure to the foetus before pregnancy is determined. These are the guidelines adopted by the US Nuclear Regulatory Commission. Even more recently, the Commission has added a recommendation that occupational exposure to an embryo/foetus should

not exceed 50 mrem per month from the time pregnancy is recognised (Paul, 1993). The older reference containing information about exposure of fertile women and foetuses, Hunt (1979), mentions exactly the same amounts as Rayburn, based on the notion that a foetus is to be treated as a member of the public involuntarily brought into controlled areas.

The biological effect of radiation is dependent on the relative radiosensitivity of the irradiated organs. For reproductive tissue, classified as moderately highly radiosensitive by Rayburn (1990), an effect of reduced fertility will be noticeable at whole-body exposures of 100 rem, and doses between 200 and 300 rem produce infertility for a period of 12 to 15 months. This irradiation leads to changes in the germ cells and can cause mutations in future offspring. Germ cell loss in males results in infertility. In a specific study of normal volunteers who underwent a single exposure (Rowley *et al.*, 1974) no abnormalities were found with exposures to less than 10 rad. Doses of 15-50 rad decreased sperm concentration over the course of three months, with recovery after 9 to 18 months. Doses over 100 rad resulted in markedly decreased sperm concentrations in 6 weeks, persisting for up to 5 years and beyond, particularly in men treated with 400 to 600 rad. The same total dose may produce more injury when delivered in divided increments (Shapiro *et al.*, 1985). To put this in perspective: some deaths occur at whole body doses of around 300 rem, although the $LD_{50/30}$ (the dose required to kill 50% of the exposed population within 30 days of exposure) for man is 450 rem. Tissues with a high turn-over of cells (*e.g.* blood-forming organs) are much more radiosensitive than very stable tissues like muscles and nerves.

For background doses, again Rayburn (1990) does not agree with the Local Rules (1986), which state an average background dose from naturally occurring sources of around 0.2 rem per year, and from average medical exposures of 0.025 rem per year. Rayburn shows an extensive table, with a total lower than that in the Local Rules (see table B.4).

Note the entry for 'television sets' in table B.4. Comprehensive information is also available on radiation emissions from Video Display Units (Marriott and

Source	Dose (mrem/year)
Medical use	
Diagnosis	77
Dental	1.4
Radiopharmaceuticals	13.6
Total	92
Natural background	
Cosmic	28
Terrestrial	26
Internal	28
Total	82
Technological development	
Fallout	5
Power plant and nuclear industry	<1
Building materials	5
Television sets	0.5
Aircraft travel	0.5
Total	12
Grand total	186

Table B.4: Total annual radiation dose from nonoccupational, environmental sources, compiled for USA population. Reproduced from Rayburn(1990)

Stuchly, 1986). Data are given about infrared, optical and ultraviolet radiation, X-ray emissions, radiofrequencies, extremely low frequencies, electrostatic fields and acoustic radiation supposedly produced by VDUs. Most of these are emitted in measurable amounts, but not large enough to produce biological effects. Especially X-ray emissions have been subject of research. The main conclusion is that, due to premarket testing, the emissions are extremely low, even below the natural background, whatever the age of the VDU. Among the nonionising radiation, only radiofrequencies receive a mention in the recommendations of the article, namely that it is good practice to avoid unnecessary exposure to radiowaves from the back

and sides of adjacent VDUs. Within one metre there is a chance of overexposure, if the VDU has not been tested explicitly and found to have low emission.

With respect to reproduction, Marriott and Stuchly devote one section to abnormal outcomes of pregnancy. Although there has been ample media coverage, detailed investigation has been difficult because of the small sample size. However, even in studies where the statistics are acceptable, the results are negative. There are no statistically significant differences in pregnancy outcomes between VDU-exposed and non-exposed women.

Although an accident with ionising radiation in a biomedical laboratory should be treated as an emergency, having a job with radiation exposure does not reduce one's average life expectancy any more than the risk of drowning (40 and 41 days, respectively) (Rayburn, 1990). In contrast, smoking cigarettes takes away 2,250 days for a male and 800 days for a female, and motor vehicle accidents 207 days. Additionally, all safety improvements made between 1966 and 1976 have increased life expectancy by 110 days.

In addition to the risks of ionising radiation from various sources and nonionising radiation from VDUs as described above, a biomedical laboratory may contain other sources of nonionising radiation like microwaves, laser, ultraviolet light and ultrasound.

Commercially available microwave ovens are tested immediately after production, and should not reach the customer if unsafe. So in practice they emit little or no stray radiation. However, they shoud be surveyed annually to ensure satisfactory performance. Microwave devices built into research apparatus should be tested during construction, regularly (quarterly) afterwards, and whenever the setup is changed (Fuscaldo *et al.*, 1980). According to Shepard (1980) microwave energy is too low for direct effects, regardless of dose; any noticeably effect is due to the heat generated. Hunt (1979), however, claims that microwaves have an effect on male fertility. It is not clear whether the reported effect could again be due to hyperthermia.

For lasers, the eye is the critical organ. Dose limits are based on numerous

parameters, *e.g.* whether the laser operates in continuous versus pulsed emission mode, and the beam divergence. Protection must be chosen accordingly, and appliances should be surveyed upon installation, then on a regular basis (annually), and whenever changes are made (Fuscaldo *et al.*, 1980). It appears the use of laser equipment does not pose particular risks to reproduction. In fact, several types of laser are actively being used to enhance the chances of pregnancy in lower fertility women (Spitzer *et al.*, 1995, Fukaya *et al.*, 1995, Kainz *et al.*, 1994) and even during gamete intrafallopian transfer to treat endometriosis (Surrey and Hill, 1994) or to manipulate the male gamete to enhance fertilisation (Obruca *et al.*, 1994). This does not mean that lasers are harmless, for there are reports of the mutagenic risks of certain UV lasers (Hendrich *et al.*, 1995) and of the, fortunately reversible, effect of Argon-lasers on the colour vision of the laser operator (Arden and Hall, 1995).

Following on from the last points on lasers, it can be seen that ultraviolet light presents a hazard to the eye and skin. However, it probably has no effect on the reproductive system (Fuscaldo *et al.*, 1980).

Ultrasound is hazardous only upon direct contact with a portion of the body (Fuscaldo *et al.*, 1980). Shepard (1980) claims the energy is too low, and any reported effects are due to hyperthermia. However, this most likely relates to incidents without direct contact. Doppler ultrasound is a familiar diagnostic technology during pregnancy. No associations can be made between the use of Doppler ultrasound in pregnancy and childhood maldevelopment, although the incidence of left-handedness and low birth weight after multiple exposures might rise significantly (Salvesen and Eiknes, 1995).

B.5.2 Temperature and other possible stressors

Under normal circumstances, the core body temperature of healthy humans seldom fluctuates more than 0.6°C and the foetal core body temperature exceeds maternal temperature by approximately 0.5°C, creating a favourable gradient for foetomaternal heat transfer (Paul, 1993). Several studies indicate that thermoregulation in pregnant women is equally or even more efficient than in non-pregnant women

(Jones *et al.*, 1985, Vaha-Eskeli *et al.*, 1991). Hyperthermia has been studied extensively in animals, and has clearly shown teratogenic effects, dependent on the degree, frequency and duration of the temperature elevation, and the specific gestational period during which the temperature elevation takes place. Most of the human data on the effects of temperature are derived from studies on febrile illness during pregnancy. Although the data suggest that sustained fevers of 38.9°C or more in early pregnancy may have a teratogenic effect on foetal development, none of the studies can make a clear distinction between effects from the fever, from the illness causing the fever, or from the treatments of either fever or illness (Paul, 1993). In view of occupational safety, Paul recommends that a job should be modified if a pregnant woman complains of discomfort while working in hot environments or when wearing personal protective equipment that hinders heat transfer. Paul also mentions that even at only mildly elevated temperatures, pregnant workers whose jobs involve prolonged standing or strenuous physical activity may experience problems, especially in the last trimester of pregnancy. A second physical stressor is involved: exertion. Prolonged standing will mean reduced cardiac output because of decreased venous return, while physical activity will cause dilation of blood vessels in the skin to dissipate heat generated by increased metabolic rate. Together, these factors may cause dizziness or even temporary loss of consciousness because of decreased blood pressure. Because of this, German law states that pregnant women (and breast-feeding mothers) may not be forced to do heavy physically demanding jobs that lead to an unbearable physical burden. In particular, their jobs should not involve the regular lifting or transport of loads of more than 5 kg or the occasional lifting or transport of loads of more than 10 kg by hand (*i.e.* without mechanical aids), the operation of machinery by foot, repeated stretching or bending, or prolonged stooping or squatting (Ministerium für Arbeit, 1992). In addition to this, Dutch law states that in jobs where a pregnant woman is exposed to vibrations, her employer must temporarily provide alternative employment upon request (Ministerie van Sociale Zaken, 1994). Colie (1993) reviews numerous studies on standing, lifting, strenuous work and constancy and length of work in pregnancy, and their effect on preterm labour. In

most of the studies the outcome is statistically weak if at all significant, although the avoidance of "occupational fatigue" by taking one rest week per month in the second half of pregnancy seems beneficial.

One other stressor deserves a mention in this section: anxiety, depression and other mental stress. Pregnancy in itself is a period full of emotions and tensions. The stress at work which a non-pregnant woman may cope with well, could become too much once she is pregnant. This can have a detrimental effect on both mother and child, and can be prevented by temporarily changing work times and work load. Perkin *et al.* (1993) have studied the effect of anxiety and depression on pregnancy outcome, and come to the conclusion that these circumstances may be undesirable, but they seem of little importance in pregnancy outcome. They can not, however, rule out the existence of negative effects on women who were confident before conception but become anxious or depressed during or as a result of pregnancy.

References

Abel, J.J. (1926) *Proc. Natl. Acad. Sci. (Wash)* **12**, 132-136

Adams, M.J., Blundell, T.L., Dodson, E.J., Dodson, G.G., Vijayan, M., Baker, E.N., Harding, M.M., Hodgkin, D.C., Rimmer, B. and Sheat, S. (1969) *Nature* **224**, 491-495

Agarwal, R.C. (1978) *Acta Cryst.* **A34**, 791-809

Albisser, A.M., Leibel, B.S., Ewart, T.G., Davidovac, Z., Botz, C.K. and Zingg, W. (1974) *Diabetes* **23**, 389-396

Albrecht, G.A. and Corey, R.B. (1939) *J. Am. Chem. Soc.* **61**, 1087

Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rodgers, J.R. and Watson, D.G. (1979) *Acta Cryst.* **B35**, 2331-2339

Arden, G.B. and Hall, M.J. (1995) Eye **9**, 686-696

Baker, E.N., Blundell, T.L., Cutfield, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G., Crowfoot Hodgkin, D.M., Hubbard, R.E., Isaacs, N.W., Reynolds, C.D., Sakabe, K., Sakabe, N. and Vijayan, N.M. (1988) *Phil. Trans. Roy. Soc. Lond.* **319**, 369-456

Banting, F.G. and Best, C.H. (1922) *J. Lab. Clin. Med.* **7**, 251-266

Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E. and Goodman, H.M. (1980) *Nature* **284**, 26-32

Beks, P.J., Mackaay, A.J.C., Neeling, J.N.D. de, Vries, H. de, Bouter, L.M. and Heine, R.J. (1995) *Diabetologia* **38**, 86-96

Bentley, G., Dodson, E., Dodson, G., Hodgkin, D. and Mercola, D. (1976) *Nature* **261**, 166-168

Bentley, G.A., Brange, J., Derewenda, Z., Dodson, E.J., Dodson, G.G., Markussen, J., Wilkinson, A.J., Wollmer, A. and Xiao, B. (1992) *J. Mol. Biol.* **228**, 1163-1176

Berg, P., Baltimore, D., Boyer, H.W., Cohen, S.N., Davis, R.W., Hogness, D.S, Nathans, D., Roblin, R., Watson, J.D., Weissman, S. and Zinder, N.D. (1974) *Science* **185**, 303

Bernal, J.D. and Crowfoot, D. (1934) *Nature* **133**, 794

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535-542

Bliss, M. (1982) The discovery of insulin (McClelland and Stewart, eds). University of Chicago Press, Chicago

Bloom, C.R., Choi, W.E., Brzović, P.S., Ha, J.J., Huang, S.T., Kaarsholm, N.C. and Dunn, M.F. (1995) *J. Mol. Biol.* **245**, 324-330

Blow, D.M. and Crick, F.H.C. (1959) *Acta Cryst.* **12**, 794-802

Blundell, T., Dodson, G., Hodgkin, D. and Mercola, D. (1972) *Adv. Protein Chem.* **26**, 279-402

Bottazzo, G.F. (1993) *Diabetes* **42**, 778-800

Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) *Science* **253**, 164-170

Brange, J. (1987) Galenics of Insulin - The Physico-chemical and Pharmaceutical Aspects of Insulin and Insulin Preparations, Springer-Verlag Berlin Heidelberg, Germany

Brange, J., Ribel, U., Hansen, J.F., Dodson, G., Hansen, M.T., Havelund, S., Melberg, S.G., Norris, F., Norris, K., Snel, L., Sørenson, A.R. and Voigt, H.O. (1988) *Nature* **333**, 679-682

Bricogne, G. (1976) *Acta Cryst.* **A32**, 832-847

Bricogne, G. (1984) *Acta Cryst.* **A40**, 410-445

Bricogne, G. (1988) *Acta Cryst.* **A44**, 517-545

Bricogne, G. and Gilmore, C.J. (1990) *Acta Cryst.* **A46**, 284-297

Bricogne, G. (1991) in: Direct Methods of Solving Crystal Structures (Schenk, ed). Plenum Press, New York. 157-175

Bricogne, G. (1991a) in: Mathematics and Molecular Biology II: DNA Sequence to Protein Structure. Santa Fe, New Mexico, USA, 24-29 March 1991, Abstract S2

Bricogne, G. (1991b) in: Second European Workshop on the Crystallography of Biological Macromolecules. Como, Italy, 13-16 May 1991, Abstract 25

Bricogne, G. (1992) in: Molecular Replacement. Proceedings of the CCP4 Study Weekend 1992, 62-75.

Bricogne, G. (1993) *Acta Cryst.* **D49**, 37-60

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) *J. Comput. Chem.* **4**, 187-217

Brown, L., Munoz, C., Siemer, L., Edelman, E., and Langer, R. (1986) *Diabetes* **35**, 692-697

Brünger, A.T., Kuriyan, K. and Karplus, M. (1987) *Science* **235**, 458-460

Brünger, A.T. (1990) *Acta Cryst.* **A46**, 46-57

Brünger, A.T. (1992) *Nature (London)* **355**, 472-474

Brünger, A.T. (1996) in: *Methods in Enzymology.* In press

Brzović, P.S., Choi, W.E., Borchardt, D., Kaarsholm, N.C. and Dunn, M.F. (1994) *Biochemistry* **33**, 13057-13069

Buchwald, H., Varco, R.L., Rupp, W.M., Goldenberg, F.J., Barbosa, J., Rohde, T.D., Schwartz, R.A., Rublein, T.G. and Blackshear, P.J. (1981) *Lancet* **I**, 1233-1234

Callahan, T., Gleason, W.B. and Lybrand, T.P. (1990) *J. Appl. Cryst.* **23**, 434-436

Castellano, E.E., Oliva, G. and Navaza, J. (1992) *J. Appl. Cryst.* **25**, 281-284

Collaborative Computational Project, Number 4 (1994) *Acta Cryst.* **D50**, 760-763

Chang, W., Stuart, D., Dai, J., Todd, R., Zhang, J., Xie, D., Kuang, B. and Liang, D. (1986) *Scientia Sinica B* **29**, 1273-1284

Chibnall, A.C. (1942) *Proc. Roy. Soc. (London)* **131B**, 136

Christianson, D.W. (1991) *Adv. Prot. Chem.* **42**, 281-355

Ciszak, E. and Smith, G.D. (1994) *Biochemistry* **33**, 1512-1517

Ciszak, E., Beals, J.M., Frank, B.H., Baker, J.C., Carter, N.D. and Smith, G.D. (1995) *Structure* **3**, 615-622

Clement Associates, Inc., Council on Environmental Quality (1981) in: Chemical Hazards to Human Reproduction, US Government Printing Office, Washington USA. Chapter II, 3

Cohen, G.H. (1993) *J. Appl. Cryst.* **26**, 495-496

Colie, C.F. (1993) *Sem. in Perinat.* **17**, 37-44

Collins, D.M. (1982) *Nature* **298**, 49-51

Colovos, C. and Yeates, T.O. (1993) *Protein Sci.* **2**, 1511-1519

Coustan, D.R. (1993) *Diabetes Care* **16**, 8-15

Cowtan, K.D. (1996) in: The Refinement of Protein Structures. Proceedings of the CCP4 Study Weekend 1996, 23-28

Crea, R., Kraszewski, A., Hirose, T. and Itakura, K. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5765-5769

Creque, H.M., Langer, R. and Folkman, J. (1980) *Diabetes* **29**, 37-40

Crowfoot, D.M. (1935) *Nature* **135**, 591-592

Crowfoot, D. (1938) *Proc. Roy. Soc.* **A164**, 580-602

Crowfoot, D.M. and Riley, D. (1939) *Nature* **144**, 1011-1012

Crowther, R.A. and Blow, D.M. (1967) *Acta Cryst.* **23**, 544-548

Crowther, R.A. (1972) in: The Molecular Replacement Method (Rossmann, ed). Gordon and Breach, New York. 173-178

Cruickshank, D.W.J. (1996) in: The Refinement of Protein Structures. Proceedings of the CCP4 Study Weekend 1996, 11-22

Cutfield, S.M. (1975) D.Phil Thesis, University of Oxford

Cutfield, J.F., Cutfield, S.M., Dodson, E.J., Dodson, G.G., Emdin, S.O. and Reynolds, C.D. (1979) *J. Mol. Biol.* **132**, 85-100

Dapergolas, G. and Gregoriadis, G. (1976) *Lancet* **II**, 824-827

Deen, J. van (1864) *Centralblatt für die medic. Wissenschaften* **2**

Derewenda, U., Derewenda, Z., Dodson, E.J., Dodson, G.G., Reynolds, C.D., Smith, G.D., Sparks, C. and Swenson, D. (1989) *Nature Lond.* **338**, 594-596

Derewenda, U. (1990) D.Phil Thesis, University of York

Derewenda, U., Derewenda, Z., Dodson, E.J., Dodson, G.G., Xiao, B. and Markussen, J. (1991) *J. Mol. Biol.* **220**, 425-433

Dodson, E.J., Dodson, G.G., Lewitova, A. and Sabesan, M. (1978) *J. Mol. Biol.* **125**, 387-396

Dodson, G., Glusker, J.P. and Sayre, D. (eds) (1981) Structural Studies on Molecules of Biological Interest - A volume in honour of Dorothy Hodgkin. Clarendon Press, Oxford

Dodson, G.G., Hubbard, R.E. and Xiao B. (1991) in: Molecular Conformation and Biological Interactions (Balaram and Ramaseshan, eds). Indian Academy of Sciences, Bangalore. 97-113

Dodson, E.J., Dodson, G.G., Hubbard, R.E., Moody, P.C.E., Turkenburg, J., Whittingham, J., Xiao B., Brange, J., Kaarsholm, N. and Thogersen, H. (1993) *Phil. Trans. R. Soc. Lond. A* **345**, 153-164

Dodson, E., Kleywegt, G.J. and Wilson, K. (1996) *Acta Cryst.* **D52**, 228-234

Drury, M.I. (1988) in: Clinical Diabetes - An Illustrated Text (Besser, Bodansky and Cudworth, eds). Gower Medical Publishing, London, England. 20.1

Ealick, S.E. (1995) in: Making the Most of your Model. Proceedings of the CCP4 Study Weekend 1995, 1-9

Emdin, S.O., Dodson, G.G., Cutfield, J.M. and Cutfield, S.M. (1980) *Diabetologia* **19**, 174-182

Engh, R.A. and Huber, R. (1991) *Acta Cryst.* **A47**, 392-400

Fletcher, R. and Reeves, C.M. (1964) *Comput. J.* **7**, 149-154

Fortelle, E. de la and Bricogne, G. (1996) in: *Methods in Enzymology.* In press

Frazão, C., Soares, C.M., Carrondo, M.A., Pohl, E., Dauter, Z., Wilson, K.S., Hervás, M., Navarro, J.A., De la Rosa, M.A. and Sheldrick, G.M. (1995) *Structure* **3**, 1159-1169

French, S. and Wilson, K. (1978) *Acta Cryst.* **A34**, 517-525

Fujinaga, M. and Read, R.J. (1987) *J. Appl. Cryst.* **20**, 517-521

Fukaya, T., Murakami, T., Tamura, M., Watanabe, T., Terada, Y. and Yajima, A. (1995) *Am. J. of Obstet. and Gynecol.* **173**, 119-125

Fuscaldo, A.A., Erlick, B.J. and Hindman, B. (eds) (1980) Laboratory Safety - Theory and Practice. Academic Press, Inc., New York USA

Gilmore, C.J., Henderson, K. and Bricogne, G. (1991) *Acta Cryst.* **A47**, 830-841

Glusker, J.P. (1991) *Adv. Prot. Chem.* **42**, 1-76

Goeddel, D.V., Kleid, D.G., Bolivar, F., Heyneker, H.L., Yansura, D.G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K. and Riggs, A.D. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 106-110

Goodfellow, J., Henrick, K. and Hubbard, R. (eds) (1989) Molecular Simulation and Protein Crystallography. Proceedings of the Joint CCP4/CCP5 Study Weekend 1989.

Goosen, M.F.A., Leung, Y.F., O'Shea, G.M., Chou, S. and Sun, A.M. (1983) *Diabetes* **32**, 478-481

Graaff, R.A.G. de, Lewit-Bentley, A. and Tolley, S.P. (1981) in: Structural Studies on Molecules of Biological Interest - A volume in honour of Dorothy Hodgkin (Dodson, Glusker and Sayre, eds). Clarendon Press, Oxford. 547-556

Grant, P.T., Coombs, T.L. and Frank, B.H. (1972) *Biochem. J. 126*, 433-440

Hagedorn, H.C., Jensen, B.N., Krarup, N.B. and Wodstrup, I. (1936) *J. Am. Med. Assoc.* **106**, 177-180

Hallas-Möller, K. (1956) *Diabetes* **5**, 7-14

Hansen, J.F. and Brange, J. (1987) *Prot. Eng.* **1**, 250

Harfenist, E.J. and Craig, L.C. (1952) *J. Am. Chem. Soc.* **74**, 3087-3089

Harker, D. (1936) *J. Chem. Phys.* **4**, 381-390

Harker, D. (1956) *Acta Cryst.* **9**, 1-9

Hendrich, C., Sudmann, J., Mommsen, J. and Siebert, W.E. (1995) *Z. Orthopadie und ihre Grenzgebiete* **133**, 520-523

Hendrickson, W.A. and Konnert, J.H. (1980) in: Computing in Crystallography (Diamond, Ramaseshan and Venkatesan, eds). Bangalore, Indian Academy of Sciences. 13.01-13.25

Hendrickson, W.A. (1985) *Methods in Enzymology* **115**, 252-270

Hill, C.P., Dauter, Z., Dodson, E.J., Dodson, G.G. and Dunn, M.F. (1991) *Biochemistry* **30**, 917-924

Hirai, S., Yashiki, T. and Mima, H. (1981) *Int. J. Pharm.* **9**, 165-172

Hodgkin, D.C. and Riley, D.P. (1968) in: Struct. Chem. Mol. Biol. (Rich and Davidson, eds). W.H. Freeman Co., San Francisco. 15-28

Holden, P.H. (1991) D.Phil Thesis, University of York

Howell, S.L., Montague, W. and Tyhurst, M. (1975) *J. Cell Sci.* **19**, 395-409

Howell, S.L., Tyhurst, M., Duvefelt, H., Andersson, A. and Hellerström, C. (1978) *Cell Tissue Res.* **188**, 107-118

Hubbard, T.J.P. and Blundell, T.L. (1987) *Prot. Eng.* **1**, 159-171

Huber, R. and Schneider, M. (1985) *J. Appl. Cryst.* **18**, 165-169

Hunt, V.R. (1979) Work and the Health of Women, CRC Press, Inc., Boca Raton, Florida USA

Jacoby, E., Hua, Q.X., Stern, A.S., Frank, B.H. and Weiss, M.A. (1996) *J. Mol. Biol.* **258**, 136-157

Jack, A. and Levitt, M. (1978) *Acta Cryst.* **A34**, 931-935

Jaynes, E.T. (1957) *Phys. Rev.* **106**, 620-630

Jaynes, E.T. (1979) in: The Maximum Entropy Formalism (Levine and Tribus, eds). Cambridge, MIT Press. 15-118

Jefferson, I.G., Marteau, T.M., Smith, M.A. and Baum, J.D. (1985) *Diabetic Medicine* **2**, 493-495

Jensen, H. and Evans, E.A. (1935) *J. Biol. Chem.* **108**, 1

Johnston, C.F., Shaw, C., O'Hare, M.M.T. and Buchanan, K.D. (1988) in: Clinical Diabetes - An Illustrated Text (Besser, Bodansky and Cudworth, eds). Gower Medical Publishing, London, England. 1.1

Jones, R.L., Botti, J.J., Anderson, W.M and Bennett, N.L. (1985) *Obstet. Gynecol.* **65**, 340-345

Jones, T.A. (1978) *J. Appl. Cryst.* **11**, 268-272

Jones, T.A., Zou, J.-Y., Cowan, S.W. and Kjeldgaard, M. (1991) *Acta Cryst.* **A47**, 110-119

Jorpes, J.E. (1949) *Arch. Intern. Med.* **83**, 363-371

Kaarsholm, N.C., Ko, H.C. and Dunn, M.F. (1989) *Biochemistry* **28**, 4427-4435

Kaarsholm, N.C., Havelund, S. and Hougaard, P. (1990) *Arch. Biochem. Biophys.* **283**, 496-502

Kainz, C., Sam, C., Sliutz, G., Kolbl, H., Dadak, C. and Reinthaller, A. (1994) *J. Gyn. Surgery* **10**, 221-227

Kinneging, A.J. (1986) Doctoral thesis, University of Leiden, The Netherlands

Konnert, J.H. and Hendrickson, W.A. (1980) *Acta Cryst.* **A36**, 344-350

Kott, A. (1996) *Countdown* **17**, 22-26

Kraulis, P.J. (1991) *J. Appl. Cryst.* **24**, 946-950

Krayenbühl, C. and Rosenberg, T. (1946) *Rep. Steno. Mem. Hosp.* **1**, 60-73

Kung, Y.T., Du,.Y.C., Huang, W.T., Chen, C.C., Ke, L.T., Hu, S.C., Jiang, R.Q., Chu, S.Q., Niu, C.I., Hsu, J.Z., Chang, W.C., Cheng, L.L., Li, H.S., Wang, Y., Loh, T.P., Chi, A.H., Li, C.H., Shi, P.T., Yieh, Y.H., Tang, K.L. and Hsing, C.Y. (1965) *Scientia sin.* **14**, 1710

Lamzin, V.S. and Wilson, K.S. (1992) in: Molecular Replacement. Proceedings of the CCP4 Study Weekend 1992, 76-83

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M (1993) *J. Appl. Cryst.* **26**, 283-291

Lee, J and Pilch, P.F. (1994) *Am. J. Physiology* **266**, C319-C334

Leslie, A.G.W., Brick, P. and Wonacott, A.J. (1986) *CCP4 Newsletter* **18**, 33-39

Levitt, M. (1974) *J. Mol. Biol.* **82**, 393-420

Liang, D., Zhuli, W., Zhang, Y. and Cao, Q. (1982) *Scientia Sinica B* **27**, 23-27

Local Rules for Persons Working with Ionising Radiations (1986) University of York

Ludvigsen, S., Roy, M., Thøgersen, H. and Kaarsholm, N.C. (1994) *Biochemistry* **33**, 7998-8006

Luzzati, V. (1952) *Acta Cryst.* **5**, 802-810

Luzzati, V. (1953) *Acta Cryst.* **6**, 142-152

Mackay, K.M. and Mackay R.A. (1981) Introduction to Modern Inorganic Chemistry **3rd ed.** International Textbook Company London

Main, P. (1979) *Acta Cryst.* **A35**, 779-785

Markussen, J. (1982) in: Neue Insuline (Petersen, Schlüter and Kerp, eds). Freiburger Graphische Betriebe, Freiburg. 38-44

Markussen, J., Damgaard, U., Diers, I., Fiil, N., Hansen, M.T., Larsen, P., Norris, F., Norris, K., Schou, O., Snel, L., Thim, L. and Voigt, H.O. (1986) *Diabetologia* **29**, A568-A569

Markussen, J., Diers, I., Engesgaard, A., Hansen, M.R., Hougaard, P., Langkjaer, L., Norris, K., Ribel, U., Sørenson, A.R., Sørenson, E. and Voigt, H.O. (1987) *Prot. Eng.* **1**, 215-223

Marriott, I.A. and Stuchly, M.A. (1986) *J. Occup. Med.* **28**, 833-848

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) *J. Chem. Phys.* **21**, 1087-1092

Meyer, J. de (1909) *Arch. Fisiol.* **7**, 96

Ministerie van Sociale Zaken en Werkgelegenheid (1994) Info. Centrale Directie Voorlichting, Bibliotheek en Documentatie, Den Haag

Ministerium für Arbeit und Soziales des Landes Sachsen-Anhalt (1992) Mutterschutz. Ziele Druck GmbH, Feldkirchen/München

Mooy, J.M., Grootenhuis, P.A., Vries, H. de, Valkenburg, H.A., Bouter, L.M., Kostense, P.J. and Heine, R.J. (1995) *Diabetes Care* **18**, 1270-1273

Murshudov, G.N., Dodson, E.J. and Vagin, A.A. (1996) in: The Refinement of Protein Structures. Proceedings of the CCP4 Study Weekend 1996, 93-104

Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1996b) In press

Navaza, J. (1987) *Acta Cryst.* **A43**, 645-653

Navaza, J. (1990) *Acta Cryst.* **A46**, 619-620

Navaza, J. (1994) *Acta Cryst.* **A50**, 157-163

Navaza, J. and Vernoslova, E. (1995) *Acta Cryst.* **A51**, 445-449

National Diabetes Data Group (1979) *Diabetes* **28**, 1039-1057

Obruca, A., Strohmer, H., Sakkas, D., Menezo, Y., Kogosowski, A., Barak, Y. and Feichtinger, W. (1994) *Human Reproduction* **9**, 1723-1726

Oldfield, T.J. (1992) *J. Mol. Graphics* **10**, 247-252

Oldfield, T.J. (1996) In preparation

Otwinowski, Z. (1991) DENZO. A Film Processing Program for Macromolecular Crystallography. Dallas, Texas, USA

Patterson, A.L. (1935) *Z. Krist.* **90**, 517-542

Paul, M.E. (1993) *Sem. in Perinat.* **17**, 5-16

Peking Insulin Structure Group (1971) *Peking Rev.* **40**, 11-16

Peking Insulin Structure Group (1974) *Scientia Sin.* **17**, 752

Peking Insulin Structure Group (1976) *Scientia Sin.* **19**, 358

Perkin, M.R., Bland, J.M., Peacock, J.L. and Anderson, H.R. (1993) *Br. J. of Obstet. Gynaecol.* **100**, 629-634

Peschar, R. and Schenk, H. (1987) *Acta Cryst.* **A43**, 513-522

Pfeiffer, E.F., Thum, C. and Clemens, A.H. (1974) *Horm. Metab. Res.* **6**, 339-342

Ponder, J.W. and Richards, F.M. (1987) *J. Mol. Biol.* **193**, 775

Quanta96 beta release, February 1996, User's Reference. San Diego: Molecular Simulations, 1996

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V.J. (1963) *J. Mol. Biol.* **7**, 95-99

Rayburn, S.R. (1990) The Foundations of Laboratory Safety - A Guide for the Biomedical Laboratory. Springer-Verlag New York Inc., USA

Read, R.J. (1986) *Acta Cryst.* **A42**, 140-149

Read, R.J. (1990) *Acta Cryst.* **A46**, 900-912

Read, R.J. (1991) in: Crystallographic Computing 5, From Chemistry to Biology (Moras, Podjarny and Thierry, eds). IUCr Oxford Science Publications. 158-167

Read, R.J. and Schierbeek, A.J. (1988) *J. Appl. Cryst.* **21**, 490-495

Remmers, P. (1994) *Libelle* **6**, 45-47

Roberts, R.M. (1989) Serendipity - Accidental Discoveries in Science. John Wiley & Sons, Inc. New York

Rossmann, M.G. (1961) *Acta Cryst.* **14**, 383-388

Rossmann, M.G. and Blow, D.M. (1962) *Acta Cryst.* **15**, 24-31

Rossmann, M.G., Ford, G.C., Watson, H.C. and Banaszak, L.J. (1972) *J. Mol. Biol.* **64**, 237-249

Rowley, M.J., Leach, D.R., Warner, G.A. and Heller, C.G. (1974) *Radiat. Res.* **59**, 665-678

Safety Handbook (1992) Department of Chemistry, University of York

Saffran, M., Kumar, G.S., Savariar, C., Burnham, J.C., Williams, F. and Neckers, D.C. (1986) *Science* **233**, 1081-1084

Salvesen, K.A. and Eiknes, S.H. (1995) *Ultrasound in Obstet. Gynecol.* **6**, 293-298

Sanger, F. (1945) *Biochem. J.* **39**, 507

Sanger, F. (1949) *Biochem. J.* **44**, 126

Sanger, F. (1949) *Biochem. J.* **45**, 563

Sanger, F. and Tuppy, H. (1951) *Biochem. J.* **49**, 463

Sanger, F. and Tuppy, H. (1951) *Biochem. J.* **49**, 481

Sanger, F. and Thompson, E.O.P. (1953) *Biochem. J.* **53**, 353, 366

Sanger, F., Thompson, E.O.P. and Kitai, R. (1955) *Biochem. J.* **59**, 509

Sanger, F. (1959) *Science* **129**, 1340-1344

Schäffer, L. (1994) *Eur. J. Biochem.* **221**, 1127-1132

Schimper, A.F.W. (1881) *Z. Krist.* **5**, 131-168

Schlichtkrull, J. (1958) Insulin crystals. Munksgaard, Copenhagen

Schlichtkrull, J. (1959) in: Diabetes Mellitus III. Kongress der International Diabetes Federation Düsseldorf (Oberdisse and Jahnke, eds). Georg Thieme. Stuttgart. 773-777

Schlichtkrull, J., Funder, J. and Munck, O. (1961) in: 4e Congrès de la Féderation internationale du Diabète, Genève (Demole, ed). Editions Médecine et Hygiène. Genève. 303-305

Schlichtkrull, J., Brange, J., Ege, H., Hallund, O., Heding, L.G., Jörgensen, K.H., Markussen, J., Stahnke, P., Sundby, F. and Vølund, Aa. (1970) *Diabetologia* **6**, 80-81

Scott, D.A. (1934) *Biochem. J.* **28**, 1592-1602

Shannon, C.E. and Weaver, W. (1949) The mathematical theory of communication. Urbana, Univ. of Illinois Press

Shapiro, E., Kinsella, T.J., Mahuch, R.W., Fraass, B.A., Glatstein, E., Rosenberg, S.A. and Sherins, R.J. (1985) *J. Clin. Oncol.* **3**, 1232-1239

Sharpey-Schafer, E. (1916) The Endocrine Organs. 2nd ed., Part 2, Ch. 49, Longmans, Green, New York 1926

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. and Sieker, L.C. (1993) *Acta Cryst.* **D49**, 18-23

Shepard, T.H. (1980) Catalog of teratogenic agents, revised edition. The Johns Hopkins University Press, Baltimore, Maryland USA

Sherwood, D. (1976) Crystals, X-rays and Proteins. Longman Group Limited, London

Shiino, D., Murata, Y., Kubo, A., Kim, Y.J., Kataoka, K., Koyama, Y., Kikuchi, A., Yokoyama, M., Sakurai, Y. and Okano, T. (1995) *Journal of Controlled Release* **37**, 269-276

Sim, G.A. (1959) *Acta Cryst.* **12**, 813-815

Sippl, M. (1993) *Proteins* **17**, 355-362

Slama, G., Hautecouverture, M., Assan, R. and Tchobroutsky, G. (1974) *Diabetes* **23**, 732-738

Smith, G.D., Swenson, D.C., Dodson, E.J., Dodson, G.G. and Reynolds, C.D. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7093-7097

Smith, G.D. and Dodson, G.G. (1992) *Biopolymers* **32**, 441-445

Smith, G.D., Ciszak, E. and Pangborn, W. (1996) *Protein Sci.* **5**. In press

Spitzer, M., Herman, J., Krumholz, B.A. and Lesser, M. (1995) *Obstet. Gynecol.* **86**, 504-508

Srinivasan, R. and Ramachandran, G.N. (1965) *Acta Cryst.* **19**, 1008-1014

Steiner, D.F. (1977) *Diabetes* **26**, 322-340

Steiner, D.F. (1981) in: Structural Studies on Molecules of Biological Interest - A volume in honour of Dorothy Hodgkin (Dodson, Glusker and Sayre, eds). Clarendon Press, Oxford. 407-419

Steiner, D.F. (1983) *Harvey Lectures* **78**, 191-228

Strobino, B.R., Kline, J. and Stein, Z. (1978) *Early Human Dev.* **1**, 371

Stuart, A. and Ord, J.K. (1987) Kendall's Advanced Theory of Statistics. Vol. 1. Distribution Theory. Charles Griffin & Company Ltd, London

Sudmeier, J.L., Bell, S.J., Storm, M.C. and Dunn, M.F. (1981) *Science* **212**, 560-562

Surrey, M.W. and Hill, D.L. (1994) *J. Am. Coll. Surgeons* **179**, 440-442

Sussman, J.L., Holbrook, S.R., Church, G.M. and Kim, S.-H. (1977) *Acta Cryst.* **A33**, 800-804

Sussman, J.L. (1985) *Methods in Enzymology* **115**, 271-303

Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S. and Richardson, D.C. (1982) *J. Mol. Biol.* **160**, 181-217

Tang, Y.C. (1981) in: Structural Studies on Molecules of Biological Interest - A volume in honour of Dorothy Hodgkin (Dodson, Glusker and Sayre, eds). Clarendon Press, Oxford. 43-44

Terwilliger, T.C. and Eisenberg, D. (1987) *Acta Cryst.* **A43**, 6-13

Tolley, S.P. (1987) D.Phil Thesis, University of York

Tong, L. and Rossmann, M.G. (1990) *Acta Cryst.* **A46**, 783-792

Tronrud, D.E., Ten Eyck, L. and Matthews, B. (1987) *Acta Cryst.* **A43**, 489-501

Tronrud, D.E. (1996) in: *Methods in Enzymology.* In press

Turkenburg, J.P. (1992) D.Phil Thesis, University of York

Vaha-Eskeli, K., Erkkola, R. and Seppanen, A., (1991) *Eur. J. Obstet. Gynecol. Reprod. Biol.* **39**, 169-174

Valuev, L.I. (1995) *Vysokomolekulyarnye Soedineniya Seriya A & Seriya B* **37**, 1960-1968

Vedani, A. and Huhta, D.W. (1990) *J. Am. Chem. Soc.* **112**, 4759-4767

Vigneaud, V. du, Miller, G.L. and Rodden, C.J. (1939) *J. Biol. Chem.* **131**, 631

Vijayan, M. (1980) *Acta Cryst.* **A36**, 295-298

Vriend, G. (1990) *J. Mol. Graphics* **8**, 52-56

Vriend, G. and Sander, C. (1993) *J. Appl. Cryst.* **26**, 47-60

Walters, D.P., Smith, P.A., Marteau, T.M., Brimblet, A. and Borthwick, L.J. (1985) *Diabetic Medicine* **2**, 496-498

Weast, R.C. (ed.) (1983-1984) CRC Handbook of Chemistry and Physics **64th ed.** CRC Press Inc., Boca Raton, Florida, USA

Webster's Seventh New Collegiate Dictionary (1965) G&C Merriam Co, Springfield, Massachusetts USA

Weis, W.I. and Brünger, A.T. (1989) in: Molecular Simulation and Protein Crystallography. Proceedings of the CCP4 Study Weekend 1989, 16-29

Whittingham, J.L., Chaudhuri, S., Dodson, E.J., Moody, P.C.E. and Dodson, G.G. (1995) *Biochemistry* **34**, 15553-15563

Wigley, F.M., Londono, J.H., Wood, S.H., Shipp, J.C. and Waldman, R.H. (1971) *Diabetes* **20**, 552-556

Wilson, A.J.C. (1950) *Acta Cryst.* **3** 397-398

Wilson, A.J.C. (ed) (1992) International Tables for Crystallography Vol. C. Mathematical, Physical and Chemical Tables. Kluwer Academic Publishers, Dordrecht, The Netherlands

Wintersteiner, O., Vigneaud, V. du and Jensen, H. (1928) *J. Pharmacol. Exp. Ther.* **32**, 397-411

Wintersteiner, O. and Abramson, H.A. (1933) *J. Biol. Chem.* **99**, 741

Wodak, S.J., Pontius, J., Vaguine, A. and Richelle, J. (1995) in: Making the Most of your Model. Proceedings of the CCP4 Study Weekend 1995, 41-51

Wollmer, A., Rannefield, B., Johansen, B.R., Hejnaes, K.R., Balschmidt, P. and Hansen, F.B. (1987) *Biol. Chem. Hoppe-Seyler* **368**, 903-911

Wollmer, A., Rannefield, B., Stahl, J. and Melberg, S.G. (1989) *Biol. Chem. Hoppe-Seyler* **370**, 1045-1053

Wood, S.P., Tickle, I.J., Blundell, T.L., Wollmer, A. and Steiner, D.F. (1978) *Arch. Biochem. Biophys.* **186**, 175-183

Xiao, B. (1990) D.Phil Thesis, University of York

Zhang, X.-J. and Matthews, B.W. (1994) *Acta Cryst.* **D50**, 675-686