Ancient Protein Identification and Mass Spectrometry Data Analysis

Yue Yang

MPhil

University of York

Biology

October 2011

ABSTRACT

The aim of this MPhil study was to develop novel models and software tools for the analysis of mass-spectrometric data from degraded and ancient proteins. On the basis of background study in ancient collagen and relevant identification approaches, problems of fossil bone collagen identification were discussed. As a solution, the database named UniColl was designed as a repository of theoretical sequences generated from the known type I collagen sequences. The principle of UniColl was to contain a large number of collagen peptide sequences which can be theoretically produced under certain chemical and mathematical algorithm, to include all the known sequence variation in each peptide.

UniColl has been established and evaluated in this work. As the result, large amounts of theoretical sequence have been generated to cover as much possible collagen sequence variations as we can get based on the known information. The practical utility and quality results of this database was tested with groups of collagen sequences identified for several unknown ancient samples.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FORMULAS

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Matthew Collins and Prof. Jim Austin, for their guidance and encouragement throughout the entire course of this research. I would also like to thank the co-supervisor of my project, Dr. Leo Caves, for his invaluable advices. Assistance from Dr. Michael Buckley on the collagen sample analysis and Dr. Peter Ashton on the database application would be highly appreciated. Supports from Mrs. Julie Knox of the departmental graduation office would be sincerely appreciated.

I would also like to gratefully acknowledge the funder of this studentship, GeneTime and BioArCh.

Finally I would like to thank all my colleagues, friends and family for the constant support from all of them.

AUTHOR’S DECLARATION

I hereby declare that all the work in this thesis is solely my own, except where attributed and cited to another author.

Chapter 1

Introduction

In this research, in order to develop novel models and software tools for the analysis of mass-spectrometric data from degraded and ancient proteins, basic background of the materials that remain in ancient samples, relevant identification approaches a considered. The approach of using a redundant sequence database 'UniColl' is then discussed as the solution to the problems.

1.1 Species Identification

Species is one of the basic units of biological classification and a taxonomic rank. A species can be defined as a group of organisms placed in one taxon genus that presumed to have the same ancestors, based on their phylogenetic similarities and capability of interbreeding (Mayr *et al.* 1953). Measures have been developed to identify species on the basis of their attributes. Traditionally species were originally observed on morphological attributes or ecological niche, while more recently, molecular phylogenetic analysis has become more widely applied. Molecular similarity has been proven useful in species diagnoses and their evolutionary relationship study (Sarich and Wilson 1967).

Molecular phylogenetic analysis focuses on the hereditary molecular similarities to obtain information on organisms' taxonomy and evolutionary relationships. Generally speaking, closely related species have high molecular similarity, while patterns of molecular dissimilarity show in distantly related organisms (Fitch and Margoliash 1967). The most common approach to define phylogenetic relationship between species is the comparison of homologous sequences for molecules using sequence alignment techniques. Two main molecules involved in phylogenic study are DNA and proteins. For protein, the amino acids sequences are generally specified from the genetic codes in DNA sequences, therefore protein molecules also contain genetic information recorded from DNA (Mount 2004).

1.2 Ancient Molecular Species Identification

Species identification and their phylogenetic relationships have been studied not only in extant, but also in extinct taxa. Due to the absence of molecular information in extinct species, usually for which hard skeletal tissues or only in cases in which non diagnostic bone fragments remain for analysis, molecular analysis become a more efficient approach to recover evolutionary relationships (Stuart 1975).

In fossil bone, which is the most commonly preserved tissue in ancient samples, there are four main types of biomolecules: carbohydrates, lipids, proteins and DNA (Abelson 1954). Protein is important for species identification is because it contains genetic information that records evolutionary relationship between species and functional characteristics of organisms. Preserved amino acids have been found in fossil samples and information in fossil peptides has been analyzed (De Jong *et al.* 1974; Westbrock *et al.* 1979).

In the diagenesis and degradation process in fossil bone, DNA is likely to be highly degraded or chemically altered (Paabo *et al*. 1989). Correspondingly, proteins are more likely to survive in degraded fossil bone because of their complex, multi-level structure, where inter- and intra-molecular bonds that stabilize the molecule at each level must break for the protein to 'unfold' and expose backbone sites for peptide hydrolysis (Schweitzer 2004). This results in the internal residues of some proteins to be almost impervious to attack over extended periods of time (Eglington and Logan 1991).

1.3 Collagen

Collagen is a rigid fibrous protein that is the principle constituent of about one third of the total protein in mammalian organisms, and the main constituent of connective tissue in animals, including tendons, cartilage, bones, teeth, skin, and blood vessels. The basic unit of the collagen triple helix is a polypeptide chain, so-called α chain, consisting of the repeating sequence(-Gly-X-Y-) n, where Gly is glycine, X and Y can be any amino acid, while X is often proline and Y is often hydroxyproline (Bachmann 2005).

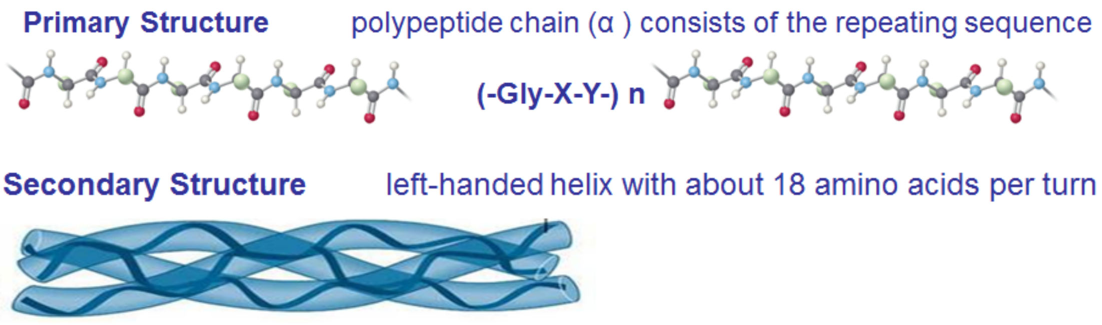

Primary Structure
polypeptide chain (α) consists of the repeating sequence
(-Gly-X-Y-) n
Secondary Structure
left-handed helix with about 18 amino acids per turn

Figure 1.3.1 - The primary and secondary strucure of collagen molecules

Three left-handed helical polypeptide units twist together, to form a right-handed triple-helical collagen molecule. Each five triple helical collagen molecules are packed side-by-side in a staggered pattern by cross-linking to form a crystalline microfibril with a 64-nm periodicity. Thousands of such microfibrils are assembled into a fibril, and then thousands of those fibrils assemble to form a fibre of tissue.

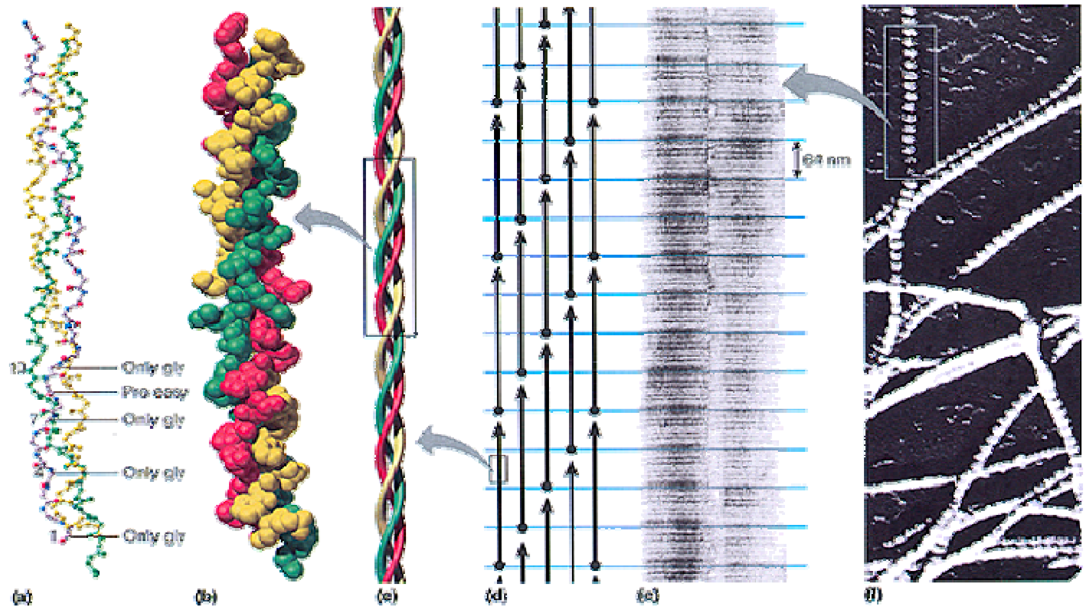


Figure 1.3.2 - Demonstration of the structure of bone collagen

The organic protein component provides flexibility and forms the matrix upon and within which mineral crystals are grown. In bone the protein phase accounts for 25-30% (by weight), of which collagen predominates accounting for about 90% (by weight) of the constituents in the organic matrix (Millard 2001). Collagens are the most abundant structural protein in the animal kingdom and of the more than 27 types of collagen, the type I collagen is prevalent, particularly so in bone.

Type I collagen is the most common form of fibrous collagen, and the major constituent of bones and skin. It comprises up to 90% of the skeletons of the mammals and is also widespread all over the body (Mirja-Liisa 2000). The molecule of most type I collagen consists of two α 1(I) and one α 2(I) polypeptide chains. Among species, α 1(I) chain is more conserved than α 2(I) chain.

1.4 Protein Sequencing Technique

The amino acid sequence of protein's polypeptide chain is a presentation of the genetic information it carries. Previously, researchers have made attempts to obtain protein sequences from laboratory, most commonly using Edman degradation method until the early 1990s. Edman degradation is a technique for protein sequencing, which relies on the identification of amino acids chemically cleaved in a stepwise fashion from the amino terminus of a peptide by reaction with phenylisothiocyanate and cleavage of the resulting phenylthiocarbamyl derivatives.

However, this method failed when the peptide being analysed possessed an acetylated or otherwise blocked amino-terminus. Condensation reactions in particular (Amadori rearrangments and/or Maillard reactions) can make the molecules increasingly insoluble and resistant to decay due to the additional formation of inter- and intramolecular cross-links (Schweitzer 2004).

Protein diagenesis can also take several forms including the conversion of one amino acid to another, loss of functional groups condensation reactions, methylation and/or glycosylation. These formations rendered such biomolecules difficult to analyse until the applications of protein mass spectrometry. Techniques using mass spectrometry (MS) overcome some of the problems associated with peptide sequencing by Edman degradation.

In recent years, mass spectrometry has been applied as one of the most common analytical techniques used to establish mass of unknown compounds, as well as a powerful tool to sequence protein molecules. Mass spectrometry is an analytical technique in which molecules from within a test sample are converted to gaseous ions (i.e. become electrically charged) that are subsequently separated in a mass spectrometer according to their mass-to-charge ratio (m/z) and detected.

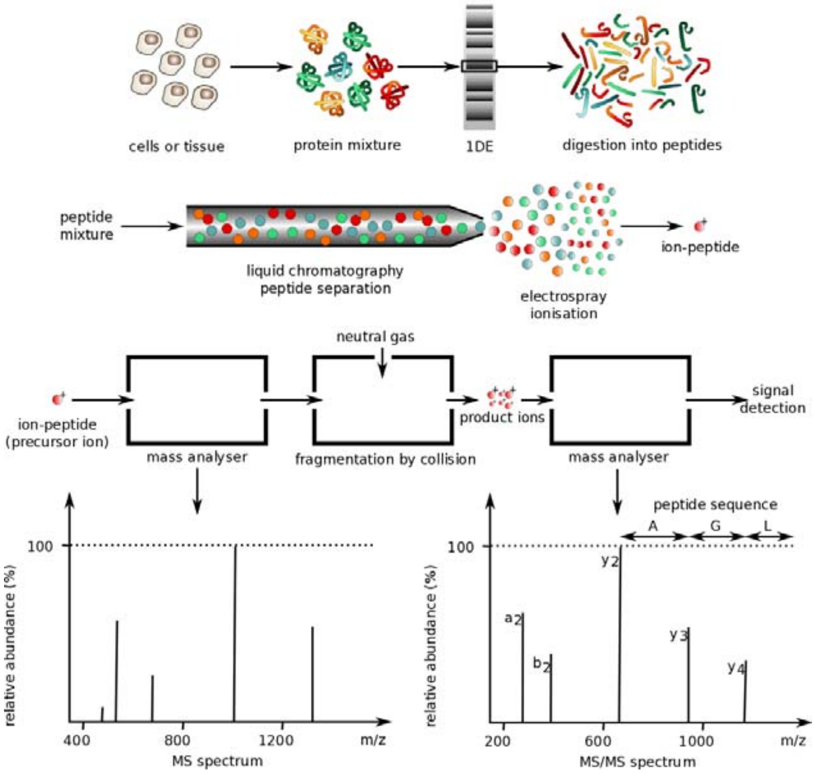


Figure 1.4.1 - Mass spectrometry process (Emmanuel Barillot et al. 2012)

Although early mass spectrometers required the sample to be in the gas phase (such as with Electron Ionisation (EI) and Chemical Ionisation (CI)), developments during the 1970s and 1980s in ionisation technologies allowed for the samples to be input as liquid solutions or solids (such as Plasma Desorption (PD), Fast Atom Bombardment (FAB) and Laser Desorption (LD)). Depending on the type of inlet and ionisation techniques used, the sample may already exist as ions in solution or it may be ionised in conjunction with its volatilisation or other methods in the ion source. 'Soft ionisation' techniques are where the evaporation and ionisation of the molecular samples into the gaseous phase are carried out without extensive fragmentation. Two of the most common ionisation methods currently used for the analysis of proteins and peptides are Matrix Assisted Laser Desorption/Ionisation (MALDI) first described by Karas and Hillenkamp (1988) and Tanaka *et al*. (1988), and Electrospray Ionisation (ESI) first described by Yamashite and Fenn (1984).

Tandem mass spectrometry (MS/MS) is a way of measuring fragment ions, most commonly generated using collision-induced dissociation (CID), for sequence interpretation. In this process, protonated molecules may be fragmented by increasing their internal energies so that they obtain sufficient energy to break internal bonds. This additional energy is most commonly transferred by collisions with a collision gas.

1.5 Protein Sequencing on the Ancient Bone Fossils

In archaeology area, the recovery of ancient proteins using MS technique provides an approach to learn more about ancient environments. A number of researchers have claimed that proteins can be found in very old bone samples, including dinosaurs by sequencing of bone extracts from protein mass spectrometry. The achievement varies from Neanderthal of 75 thousand years old (Nielsen-Marsh *et al.* 2005) and mastodon up to 300 thousand years old, to *Tyrannosaurus rex* from 68 million years ago (Asara *et al.* 2007).

Collagens, the main group of proteins in bone tissue, compose a majority of well-preserved proteins in archaeological samples. Bone collagen is arguably the most important protein for archaeologists, being used for radiocarbon dating (Bowman 1990), stable isotope analysis (Ambrose *et al.* 2003), and species identification.

The technique of sequencing protein's polypeptide chains from the mass spectrometric data is called '*De Novo*' sequencing, an algorithm used to re-generate the amino acids sequence from peptide fragments on the base of MS. In tandem mass spectrometry, the distribution of ions expresses as peaks with different mass values in the spectrum, which presents a function of the composition of the target molecule. The mechanism of '*De Novo*' sequencing is to calculate mass differences between peaks presented in the MS/MS spectrum. Synthetic examination of these mass difference values, which indicate the mass values of amino-acid components of the target peptide molecule, can lead to identifying the sequence of peptide fragments, and in the end generate the sequence of protein they composed.

Modern MS has excellent sensitivity and mass measurement accuracy. Protein sequences have been detected using mass spectrometry united with '*De Novo*' sequencing.

The process of '*De Novo*' sequencing becomes increasingly automated, while the pattern of peak masses from tandem MS can be matched against theoretical distributions of peak masses derived from peptides recorded in a database. Database searching makes protein sequencing from MS data more efficient.

However there are limitations of protein mass spectrometry database searching. Most present protein databases are species-specific. In the case of unknown species such as ancient bone, it is difficult to accurately identify its sequence using database searching, because the protein sequence which is supposed to be the correct match is not in the database.

Post-translational modifications (PTMs) further alter the protein, beyond the original DNA sequence. In the case of collagen, which is important as the key for ancient samples identification, the number of full sequences from known species is limited at the moment. As to type I collagen, which is the dominant type of collagen in bone tissue, there are few sequences in present databases. In the popular protein database 'UniProt' there were seven reliable full sequences for COL1A1 (α1 chain for type I collagen) and seven for COL1A2 (as of 1/1/2014).

The incompleteness of species coverage in protein databases induces difficulties in identifying unknown proteins by mass spectrometry database searching. In archaeological samples the problem is made worse by diagenetic damage.

1.6 The UniColl Database

Much effort has been applied to the study of fossil collagen sequences. However considering the importance of type I collagen in archaeological research and the low efficiency of MS sequencing, work need to be carried out to solve the problem of collagen sequence deficiency in current protein databases. In order to identify more ancient species from fossil bones, more collagen sequences are required to be contained in database. Therefore, novel approaches are necessarily to be developed.

The approach explored here is to generate a large collagen sequence database, containing the maximum amount of collagen sequences that can be theoretically produced under certain chemical and mathematical algorithms. This resulted in UniColl—a novel theoretical database for collagen sequence identification has been developed as a possible solution.

The UniColl database is designed as a repository of theoretical sequences generated from the known type I collagen sequences. The theoretical sequences in the UniColl are composed of all combinations of every existing variation at every position in each tryptic peptide fragment based on the alignment of type I collagen from 40 species, which were all the information could be collected from the major protein databases.

This approach is feasible on the basis of high conserved rate of amino acid composition of the collagen sequences especially for arginine (K) and lysine (R).

Because of collagen’s special molecular structure, the amino acid substitutions turned up to be much less variable compared to other types of protein. The limited sequence variation rate provides possibilities to generate theoretical sequences to include all variations from the source sequences, then to fill the deficiency of current protein databases.

The rationality behind UniColl is to explore possible solutions to problems in ancient bone collagen identification due to limitations of existing protein database, as well as to offer researchers a platform to involve novel discoveries before publication. The UniColl database provides a comprehensive data source for unknown collagen's identification, which is the major contribution of this research.

Chapter 2

Collagen Mass Spectrometry Data Analysis

In order to learn more about collagen and study the feasibility of establishing the UniColl database, some preliminary research had been carried out at the beginning of this study. The research involved similarity study in collagen sequences and MS data, noises in MS data united with the calibrating methods, and some practices of pattern recognition for collagen. The initial studies can be sorted into two main projects, which are project 1: collagen in different species; project 2: distinguishing collagen peptides.

There were eight groups of MALDI-TOF-MS data for type I collagen (M. Buckley, pers.comm, 2008) investigated in these projects. Four main species of chicken, cattle, sheep and pig were included since the type I collagen sequences for them have been fully covered in protein databases.

Table 2.1 – List of samples investigated in these projects

Index	Name	Species	Temperature
1	AI	chicken	heated at 133℃
2	AIV	chicken	heated at 145℃
3	BI	cattle	heated at 133℃
4	AUI	Auroch (ancient species close to cattle)	unheated
5	Cow1	cattle	unheated
6	Cow2	cattle	unheated
7	Sheep	sheep	unheated
8	Pig	pig	unheated

Table 2.1 listed the eight samples. 'Temperature' indicates the heating situation in sample processing. These samples will be cited as their 'Name's in the following context.

Project 1: Collagen in Different Species

The aim of project 1 is to discriminate collagen molecules come from various species. For this purpose, similarity analysis was applied on both sequences and mass spectrometry data of type I collagen. First of all, certain attributes of MS data were investigated. Then an approach of data noise filtering was developed on the original MS data in order to adjust the deviation. After that, data similarity was recomputed, and pattern recognition was applied.

2.1 Data Similarity

Although the amino acid composition of the collagen sequences are highly conserved, their mass spectrometry data are multiple and diverse. Similarity between groups of data reveals the consistent part and variable part where dissimilarity comes from. Similarity test is helpful in pattern recognition and classification, since samples with high similarity are more likely to share distinct pattern differ from others, and should be concluded in the same class.

2.1.1 Sequence Similarity

Type I Collagen sequences for different species are quite similar. As an example if we compare two samples from NCBI protein database: [gi: 109891947] (bovine type I α1 collagen); and [gi: 115268] (chicken type I α1 collagen), there is 89.5% identity between these two sequences.

Then compare the chicken sequence obtained from mass-spectrometry with the genomic sequence (from NCIB) there is a 92.4% identity. The differences are due to PTMs, mostly between hydroxyproline (B) and proline (P), while a small number differences are related to lysine (K) modification or missing, where might be the location of cross-links.

2.1.2 MS Similarity

Mass spectrometry data could be very variable. Two peptide fragments with only one difference in their sequences could generate two very different MS data. Therefore, although the difference between species exists only for a few residues in collagen sequences, this might result in quite different mass spectra. On this basis, MS similarity was tested to estimate if collagens from different species have enough dissimilarity to be classified.

In this research, the similarity between two spectra (a and b) was calculated as in Formula 2.1.1, where θ is the hypothetic spectral contrast angle that describes the difference between two spectra; i_a is the peak intensity from spectrum a, and i_b is the peak intensity from spectrum b; δ_{ab} is the m/z value difference between the two peaks. Only if both spectra have a peak at a particular m/z value, their intensities will contribute to the sum in Formula 2.1.1. In fact, if δ_{ab} is smaller than a tolerance d (setting to 0.25 in this experiment), those two peaks will be considered to have one particular m/z value. If two spectra are identical, their 'angle θ' will be 0, and $\cos\theta$ will be 1; if two spectra are completely dissimilar, θ will be right angle, and $\cos\theta$ will be 0. Here $\cos\theta$ could be regarded as the similarity between two spectra.

$$\cos\theta_{a,b} = \sum[i_a i_b (1-(\delta_{ab}/d)^2)] / \sqrt{\sum i_a{}^2 \sum i_b{}^2} \qquad \textit{Formula 2.1.1}$$

There are 180 spots in each sample each of which generate 180 spectra. Computing the similarity of each pair of spectra in one sample, and the mean of similarity of the 32400 pairs gives the similarity value of this sample as demonstrated in Formula 2.1.2, where $sim(a)$ is the similarity among spectra in sample 'a', and ai, aj are the indexes of the 180 spectra from the plate of sample 'a'.

$$sim(a) = \frac{1}{32400}\sum_{ai=1}^{180}\sum_{aj=1}^{180}\cos\theta_{ai,aj}$$

Formula 2.1.2

For two different samples 'a' and 'b', the similarity can be computed between each pair of spectra, which includes one spectrum from 'a' and another from 'b'. The average of similarity of the 32400 pairs gives the similarity value of these two samples as shown in Formula 2.1.3, where $sim(a,b)$ is the similarity between sample 'a' and 'b'; ai is the index of the 180 spectra from sample 'a'; and bj is the index of the 180 spectra from sample 'b'.

$$sim(a,b) = \frac{1}{32400}\sum_{ai=1}^{180}\sum_{bj=1}^{180}\cos\theta_{ai,bj}$$

Formula 2.1.3

Using the above algorithm, the similarity evaluation was applied on MS spectra from the eight groups of collagen samples listed in Table 2.1, and their similarity values are shown in Table 2.1.1.

As shown in Table 2.1.1, numbers on diagonal (light shading) represent the MS data similarities inside same sample set, which vary from 0.623 to 0.861, with the mean value of 72.0%. Numbers in the dark shading blocks represent the MS data similarities between same or close species, which vary from 0.608 to 0.703, with the mean value of 64.6%. Numbers in the other blocks represent the MS data similarities between different species, which vary from 0.340 to 0.613, with the mean value of 52.2%.

Table 2.1.1 - MS similarity values among the eight sample sets

	AI	AIV	BI	AUI	Cow1	Cow2	Sheep	Pig
AI	0.782	0.661	0.340	0.586	0.605	0.411	0.514	0.550
AIV		0.861	0.613	0.475	0.525	0.503	0.439	0.544
BI			0.649	0.608	0.611	0.648	0.486	0.504
AUI				0.858	0.703	0.675	0.530	0.632
Cow1					0.623	0.615	0.568	0.485
Cow2						0.638	0.603	0.459
Sheep							0.646	0.598
Pig								0.704

The light shading indicates similarity values inside the same sample set, while the dark shading indicates similarity values among the same or close species.

The distribution of similarity values in Table 2.1.1 shows that, samples from same species have higher similarities than that for different species. However the steps from those three clusters are not clear enough to distinguish different species using MS spectra by their similarity. Either the method needs improvement, or the data need amelioration; and the latter one will be discussed in the following sections.

2.1.3 Similarity-intensity Correlation

In MS data analysis, interference of small noisy peaks (commonly referred as 'grass ') adds unhelpful information from instrumental interference rather than from the sample. This causes inaccuracy in the peptide identification or sequencing results from MS data. To improve the situation, peak intensity should also be considered.

In a mass spectrum, high peak intensity is supposed to correspond to high quality of the spectrum, as well as high significance towards database matching that based on the similarity between spectra. Accordingly peak intensity and spectra similarity should be correlated. To investigate this, the relationship between the intensity and similarity for MS spectrum was tested in this work.

The intensity of a spectrum generally depends on the number of peaks and the intensity for each peak. So the sum of intensity value of all peaks in one spectrum corresponds to the intensity level of the whole spectrum. While the similarity value for each spectrum was accounted as the mean of similarity values between this spectrum and every other spectrum on the same plate as in Formula 2.1.4, where $sim(i)$ is the similarity value of the ith spot, while j is the index (1 to 180) of spots on the plate.

$$sim(i) = \frac{1}{179} \sum_{j=1, j \neq i}^{180} \cos \theta_{i,j} \qquad \textit{Formula 2.1.4}$$

After calculating the intensity and similarity for each spectrum from the 180 spots on a plate, their intensity-similarity relationship can be plotted (Figure 2.1.1 shows an example: plot of sample AI)

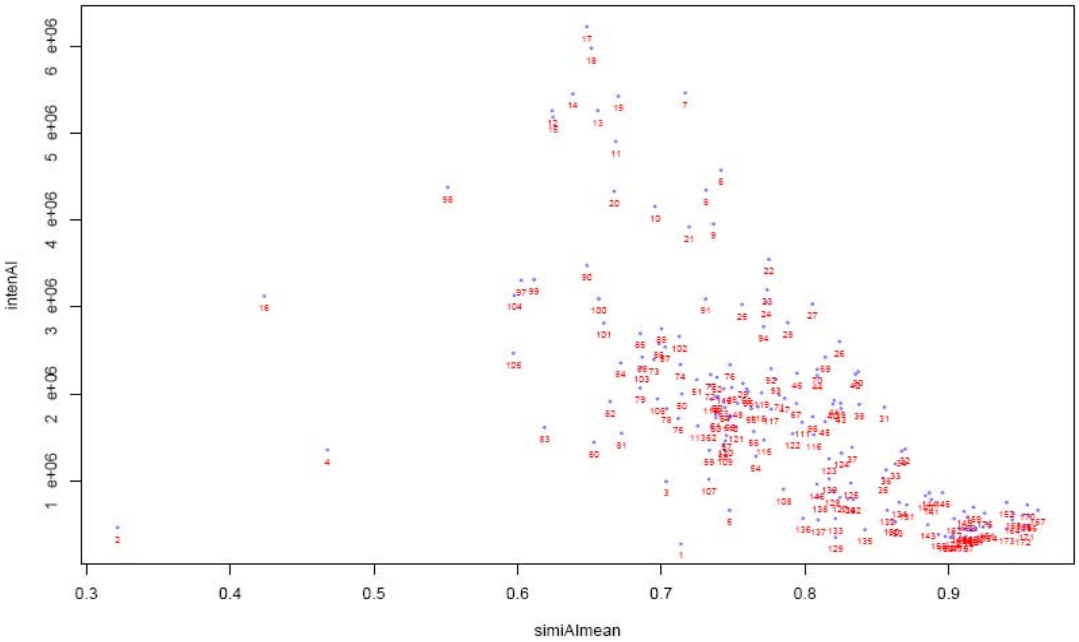


Figure.2.1.1 - MS spectra intensity and similarity correlation for AI

The 180 spots in sample set AI are plotted as 180 dots; the numbers (1 to 180) labeled under each dot indicate the spot's index. The 'simiAImean' value (x-axis) shows the mean similarity between one spot to each other; the 'intenAI' value (y-axis) shows the total intensity in the MS spectrum for that spot.

The result of the similarity-intensity correlation study for the eight samples in Table 2.1 showed that, the intensity and similarity of MS data seemed not in direct proportion as expected. There were clusters of dots in the lower right corner showed high similarity but low intensity; while most of them were labeled lager than 100. These attributions suggested the possibility of intensive matrix peaks in those spots, because they are highly identical in mass value while low in peak intensity. Matrix peak, the possible reason of the uncorrelated similarity-intensity relationship will be discussed in the coming sections.

2.2 Noises and Modification

2.2.1 Matrix Peaks

The mechanism of Matrix Assisted Laser Desorption Ionization-Time of Flight-Mass Spectrometry (MALDI-TOF-MS) is to use laser light to ablate entire polymer molecules, from a target surface into a time-of-flight mass spectrometer. A UV-absorbing organic matrix is used to facilitate the ablation of intact polymer molecules. In MALDI sample preparation for peptide and protein analysis, it's important to apply a matrix material together with the analyte to the sample support surface. So matrix peaks, with the mass lower than 800 Da, are common in mass spectra (Zaima *et al.* 2010).

Although matrix were added evenly to each of the 180 spot on a plate, the spots with only matrix peaks would appear at the end of the plate, especially after spots No.120. That might be because peptides eluted in the earlier spots during sample preparation, leaving only matrix in subsequent spots. Accordingly the late spots on the plate get few protein molecules, so their mass spectra give peaks mainly from matrix.

In the similarity-intensity plot displayed above, a group of spots appear at the right bottom with high similarity but low intensity, and most of their index numbers are larger than 100. Investigating into MS data for this group of spots reveals that, compared to spots with high-intensity values, peaks mapped on mass spectra from the low-intensity group are much fewer in numbers and smaller in mass values. Most of the low-intensity spots share similar pattern of mass peaks, which could be the reason of their high similarity values. Combined with the big index numbers that indicate to the late location, these high-similarity-and-low-intensity spots are very likely to have only matrix material but few protein residues loaded.

2.2.2 Removing Matrix Peaks

Since the existing of matrix peaks, the similarity-intensity relationship demonstrated above would be unreliable. Some spectra appear highly similar to others only because they have matrix peaks in common. The high similarity generated from these matrix peaks gives little information for protein identification. So it's necessary to filter them out.

Considering spectra that are mainly composed of matrix peaks have low intensity-similarity proportion, a cluster of this kind of spectra called 'matrix spectra' has been created, with the intensity-similarity ratio lower than 10^6. And the frequently present peaks in this cluster could be regarded as matrix peaks. Considering that matrix peaks should be lower than 800 Da, peaks bigger than that threshold were eliminated.

Since peak mass for certain fragment varies against different spot sets, the mass tolerance was specified to 0.2 Da in matrix peaks recognition. Therefore peaks presented in 'matrix spectra' with mass difference less than 0.2 Da were sorted into one matrix peak group. The number of peaks in each group indicated the frequency they presented in spectra. In order to avoid noise peaks, those groups with number of peaks less than 10% of the number of investigated 'matrix spectra' was considered less likely to be matrix.

As the result, the peak mass (in Da, after rounding) of 617, 618, 619, 620, 621, 622, 624, 626, 628, 634, 640, 642, 643, 644, 650, 651, 655, 656, 658, 660, 664, 665, 666, 668, 672, 673, 674, 676, 678, 679, 680, 681, 682, 684, 685, 688, 693, 694, 698, 700, 701, 704, 756, 758, 775, 782, 793, 794 were identified as matrix peaks that should be filtered out from all spectra before analysis in this experiment.

However these peaks were theoretical matrix peaks produced in this analysis, while some small peptide fragments might be included, and some uncommon matrix peak might be ignored. For further precision of this analysis, in future work, mass spectrometric experiments on matrix component need to be applied to decide matrix peak.

2.2.3 Peak Shift

Among the 180 spectra from one sample, peak mass values change through spots for the same peptide. For example, in the first spectrum there is a peak 666.12; in the second spectrum there is a peak 666.01; and in the third spectrum there is a peak 666.22; but they should be the same peak and indicate the same peptide fragment. However the tolerance of peak shift could be larger than 0.2 Delta, which is big enough to cause errors in data analysis. For better accuracy in the results, it was necessary to normalize the data by clearing the error, and evaluate the accuracy using similarity test again.

In order to see how peaks shifted, the peak-mass curves for several common peaks in most spots were plotted (such as ‘peak 622’ shown in Figure 2.2.1); matrix peaks are the best choice as they are present in every sample. It was found that peaks in one spectrum shifted in the same pattern, in which the mass value of a same peak varies up and down regularly through different spectra, periodically by every 24 spots. This pattern might come from the way of spots alignment on a plate, which is a 12 by 15 matrix. The mass spectrometry machine moves by rows on a plate, so we can imagine it moves from the left end of the first row to the right, and back from the right end of the second row to the left, then the same thing start with the third row. That gives a repeating cycle of peak shift with a period of 24 spots.

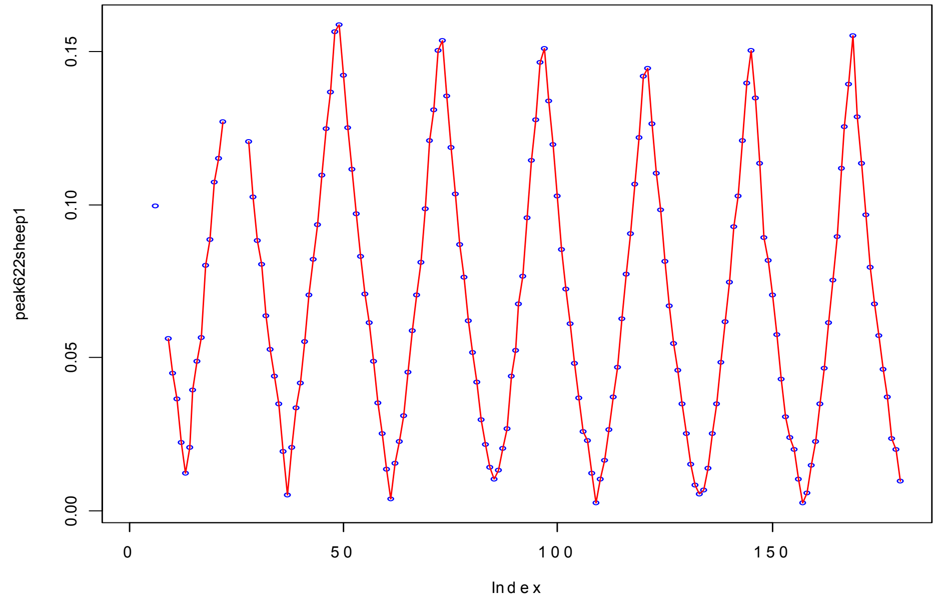


Figure 2.2.1- mass shift of peak 622 in 'Sheep'

The above figure shows how peak 622 (Da) drifting through the 180 spots in sample 'Sheep'; the x-axis is the spot index, and y-axis is the difference from experimental peak mass to the standard 622 Da.

Such periodical waveforms were discovered for the peak shift through all 180 spots on a plat, the tolerance between the peak crest and trough increases with peak mass, approximately from less than 0.1 Delta to more than 0.8 Delta, where larger peaks shift more than smaller ones. When plotted that tolerance for all peaks, they were found to be linearly correlated after linear regression (Figure 2.2.2); peaks with larger mass value generated bigger drift.

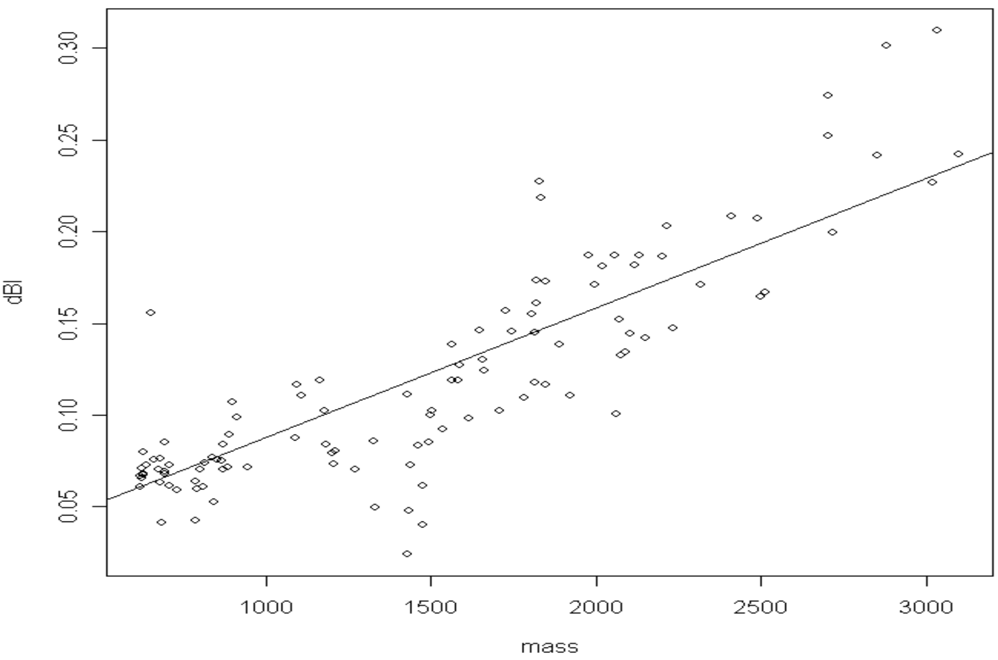


Figure 2.2.2 – peak mass-shift correlation for 'Sheep'

The above figure shows the correlation of peak shift and their mass values in sample 'Sheep'; the x-axis is peak mass (Da), and y-axis is the amount of their shift (Da).

2.2.4 Peak Mass Adjustment

The peak mass-shift correlation was analyzed and the result showed that peaks in the same spectrum shift in the same direction, while the magnitude of the shift was in direct proportion to the peak mass value. In order to adjust the peak mass value, the most common matrix peak was used as a standard in one sample. By computing the mean mass value of this matrix peak and comparing to its peak mass in all spectra, the peak shift magnitude to the mean can be calculated for every spectrum. Then for other peaks in each spectrum, the shift range can be estimated based on their peak mass as demonstrated in Formula 2.2.2.

$$\frac{mainpeakmean - mainpeakmass}{peakmean - peakmass} = \frac{t * mainpeakmass + b}{t * peakmass + b} \qquad \textit{Formula 2.2.1}$$

$$peakmean = \frac{t * peakmass + b}{t * mainpeakmass + b}(mainpeakmean - mainpeakmass) + peakmass$$

Formula 2.2.2

In the above formulas, 'peakmean' is the target mass value a peak would be adjusted to, 'peakmass' is the peak mass value read from mass spectra, 'mainpeak' is the most common peak used as a standard, 't' is the slope of the linear regression of the peak mass-shift correlation, while 'b' is the intercept.

Formula 2.2.2 is deduced from Formula 2.2.1 in order to calculate the 'peakmean'. By replacing the mass values for all peaks with 'peakmean', they would be moved back to their theoretical mean. As a result, peak mass value could be adjusted, with the tolerance varies from 0 to 0.1 Delta, which is great improvement compared to the original tolerance.

Evaluating the adjustment of sample 'AI' (type I collagen of chicken), before being adjusted, there were 20% peaks drifted over 0.05 Da, which decreased to only 8% after adjustment. The total shift distance also dropped from 30.9 Da to 17.0 Da. Other samples got similar improvement after adjustment.

Take sample 'BI' (type I collagen of cattle) for example, the most common peak 679 (Da) in its spectra was used as the standard to adjust all the other peaks. As the result, peak mass values are ideally adjusted to their mean. For instance, comparing the mass shift for peak 1095 (Da) in BI before (Figure.2.2.3) and after (Figure.2.2.4) adjustment, the tolerance reduced from more than 0.1 Da to less than 0.03 Da. And most other peaks also had good results after adjustment.

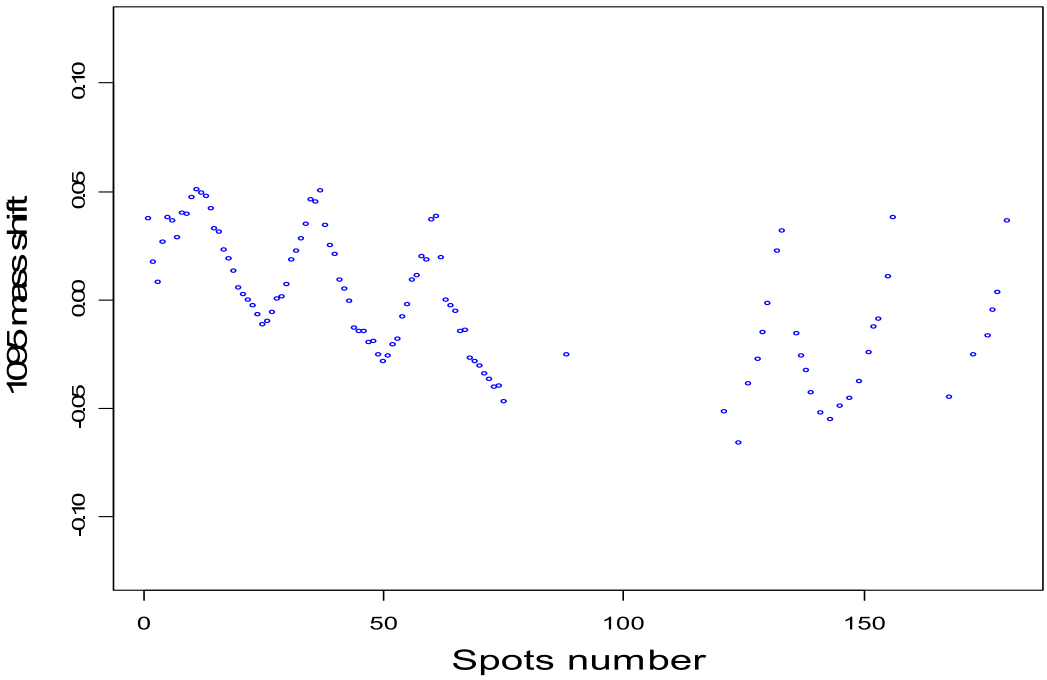


Figure 2.2.3- mass shift of peak 1095 in 'BI' BEFORE adjustment

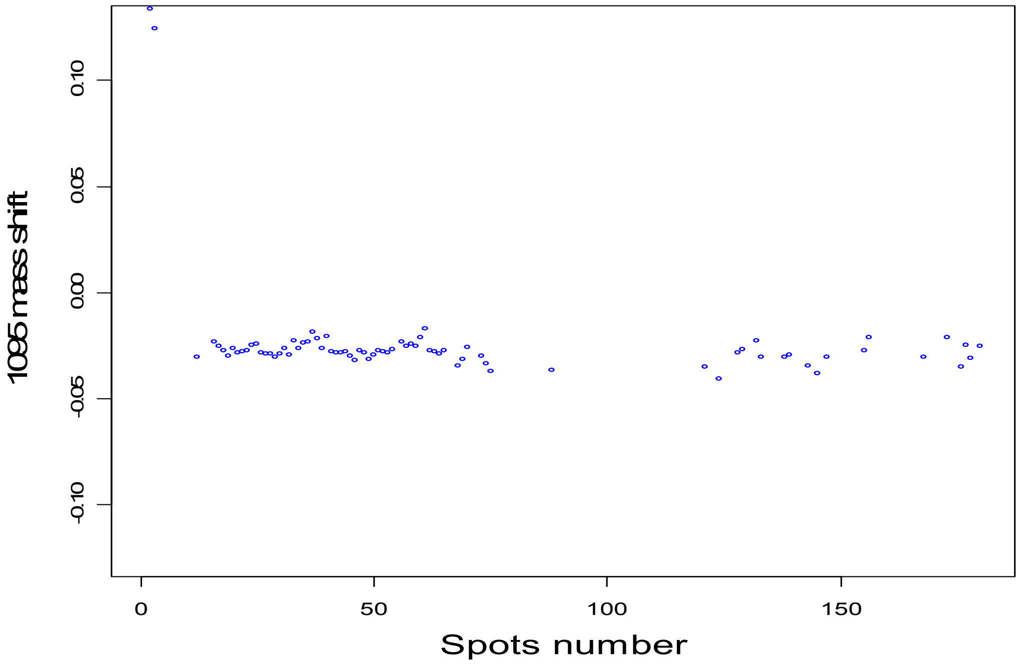


Figure 2.2.4- mass shift of peak 1095 in 'BI' AFTER adjustment

Figure 2.2.3 shows how peak 1095 (Da) shifting in sample 'BI' before data adjustment, while Figure 2.2.4 shows the situation after the adjustment. The x-axis is the spot index, and y-axis is the difference from experimental peak mass to the standard1095 Da.

The peak mass adjustment contributed more precise range for mass spectrometry data analysis, especially in similarity test. In fact, instead of setting the tolerance as 0.25 Da before, now peaks vary within 0.1 Da could be identified as similar, so that numbers of distracters with differences between 0.1 and 0.25 can be excluded.

2.2.5 MS Similarity for Normalized Data

After filtering out matrix peaks and adjusting peak mass shift, the mass spectrum data were well normalized, with more information, less noise and better accuracy. They are hereafter called normalized data. The MS similarity values were calculated for the eight sample set in Table 2.1 after data normalization, and the result is shown in Table 2.2.1.

As shown in Table 2.2.1, there comes out an obvious hierarchy of the similarity values. The first layer (diagonal) indicates the within sample similarity that is the highest; the second one (slash) is between same or closely related species; while the last layer (other) is similarity between different species that is the lowest. The similarity within same sample became approximately ten times higher than the similarity between different species.

In the similarity values before normalization (Table 2.1.1), the mean of similarities inside same sample was 72.0%, the mean of similarities in same species was 64.6%, while the mean of similarities in different species was 52.2%. The ratio for them was 1.38:1.24:1. However in the Table 2.2.1, the mean of similarities inside same sample is 39.2%, the mean of similarities in same species is 26.8%, and the mean of similarities in different species is 18.8%. The ratio for them is 2.09:1.43:1. Accordingly, after data normalization, MS presented more significant differences between different samples.

The similarity values were higher before normalization because matrix peaks made great contributions, however those were interferences that made even different sample seem similar. As the noise being filtered out, similarity became lower in values but more discriminable and reliable.

Table 2.2.1- MS similarity values among the eight sample sets after data normalizaiton

Similarity	AI	AIV	BI	AUI	Cow1	Cow2	Sheep	Pig
AI	0.465	0.180	0.047	0.119	0.249	0.242	0.290	0.224
AIV		0.373	0.085	0.219	0.141	0.175	0.110	0.119
BI			0.530	0.279	0.272	0.349	0.157	0.183
AUI				0.385	0.228	0.275	0.158	0.180
Cow1					0.286	0.296	0.262	0.243
Cow2						0.328	0.244	0.264
Sheep							0.360	0.235
Pig								0.407

The light shading indicates similarity values inside the same sample set, while the dark shading indicates similarity values among the same or close species.

2.3 Pattern Recognition

The research of sequence similarity in *Section 2.1.1* showed that, type I collagen sequences for different species are highly similar, so are their MS spectra. The common parts in protein sequences express the common peptide fragments giving common peaks in MS spectra. Therefore, a group of protein samples with high sequence similarity would share numbers of common MS peaks, which generate certain patterns to make their MS spectra look similar. In the case of type I collagen, these common peaks from MS spectra could be the marker to distinguish them from other types of protein. Although the MS spectra for type I collagen are similar, beside the common part, there always exist some peaks that are unique for certain species. These unique peaks could be the marker for the corresponding species. A pattern recognition work is needed to recognize these common peaks and unique peaks, in order to discover the peak markers to identify collagen and to distinguish species.

2.3.1 MS Peak Marker

In order to find peak markers, a clustering method has been applied. Firstly, a list of mass values of all peaks appeared in one spectrum has been generated (due to instrumental error, peaks with mass differences less than 0.2Da were considered as the same peak). Then the occurrence frequency of each peak in the list was recorded through all the spectra. At last, peaks were clustered into groups according to their occurrence frequency numbers. In comparing a pair of samples, if a peak appeared more than 20 times in both samples, it was set as a common peak for them. If one peak appeared 40 times more in one sample than in the other, it was set as a unique peak for the former sample.

Investigating two samples ‘AI’ and ‘BI’ for example, there are several common peaks in their MS spectra as shown in Table 2.3.1.

Table 2.3.1-Common peaks in ‘AI’ and ‘BI’

Common peaks in AI and BI	N Obs in AI	N Obs in BI
1976 (Da)	64	59
2057 (Da)	36	56
2316 (Da)	22	30

(N Obs= Number of observations.)

In Table 2.3.1, the first column indicates peak mass values (after rounding) of the common peaks in MS data of ‘AI’ and ‘BI’; while the other two columns shows the existing times for each common peak in the two sample sets.

For each sample, the top ten unique peaks are listed as follows (Table 2.3.2 and Table 2.3.3). These significant peaks could be used as marker for identify species in the future.

Table 2.3.2 -Unique peaks in 'AI' MS

Unique peaks in AI (Da)	N Obs in AI more than in BI
1163	67
1460	63
1573	62
1320	53
1096	51
1595	51
1464	49
1175	47
1394	47
1098	41

Table 2.3.3-Unique peaks in 'BI' MS

Unique peaks in BI (Da)	N Obs in BI more than in AI
1832	120
1648	101
1095	95
1105	90
1427	74
1562	68
1161	68
1459	64
1177	63
1328	62

(N Obs= Number of observations.)

In Table 2.3.2 and Table 2.3.3, the first columns indicate peak mass values (after rounding) of the unique peaks in MS data of 'AI' or 'BI'; while the second columns indicate how many times each unique peak exists in that sample more than in the other one.

2.3.2 MS/MS Peak Marker

Besides MS spectra, it's also necessary to look at tandem mass spectra (LC-MS/MS), because they include shorter chains (which are obscured by matrix peaks, and therefore ignored in MALDI MS), they could give more information of small fragments which are common in collagen.

Also take 'AI' and 'BI' for example, the top 10 common peaks in MS/MS spectra for each and both of them are shown in Table 2.3.4, Table2.3.5, and Table2.3.6. Obviously they share most of the common peaks. While Table 2.3.7 shows the amino acids corresponding to those common peaks.

Table 2.3.4

Common both	N obs
70	397
86	362
175	297
112	269
171	230
155	224
127	191
268	133
129	107
272	100

Table 2.3.5

Common 'AI'	N obs
70	225
86	200
175	159
112	139
171	117
155	90
127	75
268	71
272	57
283	44

Table 2.3.6

Common 'BI'	N obs
70	172
86	162
175	138
155	134
112	130
127	116
171	113
129	88
84	83
226	69

(N Obs= Number of observations.)

Table 2.3.4 shows the common peaks observed in MS/ MS data of both 'AI' and 'BI'; Table 2.3.5 and Table 2.3.6 shows the common peaks observed in MS/ MS data of 'AI' or 'BI' separately. The first columns indicate peak mass values (after rounding) of the common peaks (Da); while the second columns indicate the number of observations of each peak.

Table 2.3.7- amino acids corresponding to the common peaks

Peak mass (Da)	Corresponding amino acids
70	P
86	O
175	A (y1)
171	GO, OG (b)
155	GP, PG (b)
127	GP, PG (a)
268	GPP,PGP,PPG (b)

(P: Proline, O: Hydroxy-Proline, G: Glycine, A: Arginine)

The first column indicates peak mass values (after rounding) of the common peaks observed in MS/ MS data of sample 'AI' and 'BI'; while the second column shows the amino acids corresponding to each mass value, where the 'y1', 'a', 'b' indicate the ionizing types.

Glycine presents almost every three amino acids in collagen, proline makes up about 9% of collagen, and hydroxyl-proline derived from proline, they are the most popular amino acids in collagen peptides. Therefore the mass peaks of GPP, GP, GPP, etc. are quite common in MS/MS spectra of collagen. These peaks should be the marker of collagen. And the frequency of such fragments could be used to test the position of hydroxy-proline and cleavage.

Project 2: Distinguishing Collagen

The aim of project 2 was to distinguish collagen molecules from other types of proteins through mass spectrometric data. Most of samples involved in this research were mainly composed of collagen, but this was mixed with other proteins in tissue, or has been contaminated in earth and laboratory. In collagen identification, a principal mission is to distinguish collagen from non-collagen fragments. In this project, some attempts have been made to identify collagen using their sequence markers and hydroxylation patterns.

2.4 Collagen Marker

Collagen sequences are special with repeating units (Gly-X-Y), because glycine is a key to maintain the triple helix structure. In the repeating sequence, X and Y can be any amino acid, while X is often proline (P) and Y is often hydroxyproline (O). Accordingly, fragments composed of ‘G’, ‘P’ and ‘O’ should be commonly repeated in collagen sequences. Such internal fragments could be unique in collagen molecules, and may be the markers to distinguish collagen from other proteins.

2.4.1 Exploring Internal Fragments

In the above study in *Section 2.3*, pattern recognition approaches have been applied to find peak markers in MS and MS/MS data in order to identify samples. Here same method has been used to explore the information of collagen-specific internal fragments from tandem MS data.

To validate the existence of unique internal fragments in collagen and further detect the mass value of them, a data set was created to include all internal fragments in ten selected high quality spot sets of bone collagen and the other ten of non-collagen specimens. Their spectra were processed and analyzed by GPS (the Global Protein Server Workstation, Applied Biosystems), which uses internal Mascot software (version 2.1; Matrix Science) as the searching tool.

Mascot is a powerful searching engine developed by the Matrix Science (Perkins *et al.* 1999) to identify proteins in selected protein databases from their peptide mass fingerprints and MS/MS data. The GPS interpreted tandem MS finger print for the selected peptide, with list of sequences of ion fragments get matched in that MS/MS spectrum. These sequences were collected into an internal fragments database that was built for this analysis.

The length of these internal fragments varies from one to over twenty, while shorter fragments were observed much more frequently than longer ones. Some of the short fragments are frequent in collagen samples but rare in non-collagen. Table 2.4.1 gives a list of these collagen-specific fragments observed in the eight sample sets (Table 2.1) and their peak mass (Da). Among them, GP (a) 127, GO (a) 143, GP (b) 155, GO (b) 171, PO (a) 183, PO (b) 211, GPP (a) 224, GPO (a) 240, GPP (b) 252, GPO (b) 268 are the TOP 10 popular fragments with the highest numbers of observation.

Table 2.4.1- Internal fragments frequently observed in collagen

Sequences	a ion peak mass (Da)	b ion peak mass (Da)
GP	127	155
GO	143	171
PO	183	211
GPP	224	252
GPO	240	268
GPPG	281	309
GPOG	297	325
PPGP	321	349
POGP	337	365

(P: Proline, O: Hydroxy-Proline, G: Glycine)

The first column indicates sequences of the internal fragements; while the second and third columns show the 'a ion' or 'b ion' peak mass values (after rounding) of the corresponding sequence.

2.4.2 Verifying Collagen Markers

The collagen-specific internal fragments demonstrated above were supposed to be collagen markers. To verify the validity of the markers, an examination has been carried out on Mascot data.

Ten groups of tandem MS data of bone collagen (M. Buckley, pers.comm, 2008) were investigated in this study (Table 2.4.2). Except the five fresh samples from Table 2.1, five ancient samples were added. All of these ten samples were not heated during laboratory treatment.

Table 2.4.2 - List of samples investigated

Index	Name	Species	Fresh or Ancient
1	A4	chicken	Fresh
2	Cow1	cattle	Fresh
3	Cow2	cattle	Fresh
4	Sheep	sheep	Fresh
5	Pig	pig	Fresh
6	GT	giant tortoise	Ancient
7	Dodo	dodo	Ancient
8	MA	mammoth	Ancient
9	nrl4	mammoth	Ancient
10	ns3	mammoth	Ancient

Table 2.4.2 listed the ten samples, the first five ones in which are fresh species, while the rest ones are ancient. These samples will be cited as their 'Name's in the following context.

By searching for the collagen markers through the MS/MS data sets, most of the ten samples displayed good collagen and non-collagen discriminations, with collagen markers shown in collagen fragments rather than the non-collagen ones. Here the collagen and non-collagen discrimination was verified by Mascot, which provides reports of searching results, with information of matched proteins and peptides contained.

The result came out with lists of peptide mass values and the numbers of internal fragments (see Table 2.4.1) detected in MS/MS for each peptide (an example for nrl4 is shown in Table 2.4.3). For example, in sample dodo there were 80% collagen peptides containing those markers, while none of the non-collagen ones included them. In sample A4, collagen markers exist in 31% collagen peptides but in only 0.04% non-collagen fragments. The results suggested that these internal fragments were collagen specific, and could be used as collagen markers.

Table 2.4.3- Numbers of internal fragments detected in MS/MS for example 'nrl4'

PEAKMASS	INTERFRA	-64	-48	-32	-16	0	16	32	48	64
902.9058	0	0	1498	1503	5863	24549.773	15302	0	0	0
905.7108	0	0	0	0	0	624926.674	0	0	0	0
905.7461	0	0	0	0	0	624926.674	0	0	0	0
1018.7953	0	0	0	0	0	68026.398	0	0	0	0
1095.655	0	0	0	0	0	55501.397	0	0	0	0
1153.4955	4	0	10081	0	6067	114164.792	0	0	1482.7	0
1154.5376	4	0	26086	1203	0	657976.03	1483.5	0	64239	0
1162.6356	0	0	0	0	0	15521.332	1685.4	0	0	0
1180.5227	3	0	0	0	0	18386.444	0	0	0	0
1202.5514	3	0	6E+05	1483	0	64238.712	0	0	6226.9	0
1205.5731	7	0	0	0	0	28571.941	0	0	0	0
1208.4548	6	0	0	10017	1580	10347.849	0	0	0	0
1276.5992	4	0	0	0	0	9624.163	0	0	0	0
1328.6807	1	0	3653	0	10329	46786.38	3852.7	0	0	0
1453.8612	1	0	0	0	0	80894.333	0	0	0	3772.6
1465.6163	5	0	0	0	1207	76490.672	0	0	1189.9	41173
1467.8525	0	0	0	0	0	5731.481	0	0	0	0
1470.7642	1	0	0	0	0	16810.045	0	0	0	0

Table 2.4.3 shows internal collagen-like fragments and hydroxylation degree (delete) for peaks from 902.9058 to 1470.7642 in sample set 'nrl4'. The column 'INTERFRA' shows the number of collagen-like internal fragments in the peptide, and the following columns show the intensity of peaks with multiple of 16 differences from the peptide, indicates the probable existence of hydroxylations.

2.4.3 Limitations

Collagen markers are supposed to be detected rarely in non-collagen sequences, however practically, these peaks present in some of the non-collagen fragments, including osteocalcin, myosin, actin, serum albumin, etc, coming with the original samples or from contamination.

All proteins are composed of twenty amino acids and many combinations are possible. However collagen markers should be special for collagen in respect of the unique repeating units of (G-X-Y) and the abundance of Pro and Hyp. Although these units mainly present in collagen, other combinations of amino acids can produce equivalent mass values. However, these internal fragments are observed much less common in non-collagen sequences than collagen ones. Therefore, the difference between collagen and non-collagen should not only be defined as the existence of collagen markers, but also the frequency of the presence.

The frequency of a collagen marker appears in the MS/MS peak list depends on the quality of spectrum, related to the intensity of peaks. In high quality spectra, collagen markers are detected in non-collagen peaks. Meanwhile, in poor quality spectra, a number of collagen fragments contain no collagen markers in their spectra. Therefore the collagen-marker methodology is difficult to apply without a criterion of data quality.

2.4.4 Applications

The limitation discussed above was not only caused by the incompletion of methodology, but also the inaccuracy of Mascot searching. On the other hand, this limitation could be utilized to find out the inaccuracy in Mascot results, and look for alternative answers.

The ‘non-collagen’ m/z mentioned before are fragments that identified by Mascot as proteins other than collagen, or even non-protein residues like DNA. However Mascot is not absolutely accurate, it occasionally misidentified some residues. Especially for ancient samples, due to loss of information and lack of matches in database, they are frequently misidentified.

The collagen-marker methodology provides a way to examine the validity of Mascot searching results. Numbers of fragments that Mascot identified as non-collagen sequences were discovered to contain collagen markers with high frequency of the presence. They are probably misidentified collagens that would need to be retested by other tools such as *De Novo* sequencing.

For example, peptide with peak mass of 954.333 Da from the MS spectrum of sample set "MA" was preferentially identified as non-collagen 'GPLAPDAQGK + Deamidation (NQ)' by Mascot. However seven collagen markers were detected in the MS/MS spectrum of this peptide. Then *De Novo* sequencing was applied to detect alternative answers. As shown in *Figure 2.4.1*, the result demonstrated that this peptide could be sequenced as collagen fragment 'PLGPAGETGR', with eight of ten y ions hunted in the spectrum. It was proved that the collagen-marker methodology provided a novel pathway to discover potential collagen fragments which might be misidentiried or ignored by MS database searching.

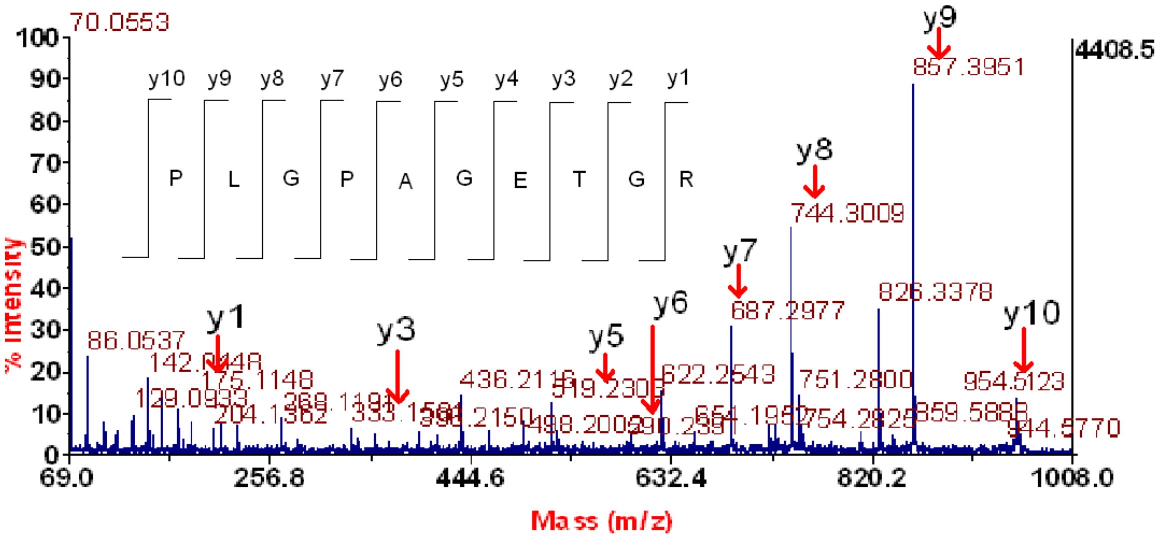


Figure.2.4.1- De Novo sequencing for precursor 954.333

The figure shows a process of *De Novo* sequencing, where the positions and mass values of the identified y ion peaks of 'y1', 'y3', and 'y5' to 'y10' are labeled.

2.5 Hydroxylation in Collagen Sequences

After collagen synthesis, hydroxyl (-OH) groups are added to some amino acids, usually proline, some times glycine and others. Hydroxylation is one of the most familiar PTMs especially in collagen.

2.5.1 Errors Caused by PTMs

Mascot database searching algorithms compress the number of spectra to be matched with sample according to peak mass. This method is efficient to reduce counting load and improve the accuracy by excluding irrespective spectra out of certain range of mass.

PTMs (post translational modifications) shift mass of peptides and expand the searching area. This will aggravate the inaccuracy of database searching by including more potential combinations of amino acids in the search space.

PTMs in collagen mainly include hydroxylation, deamidation, glycosylation and glycation. In hydroxylation, hydroxyl (-OH) groups are added to the "Y" amino acid. Glycosylation is a process of adding saccharides to collagen. It is an enzyme-directed site-specific process, as opposed to the non-enzymatic chemical reaction of glycation, which often happens in aging and ancient tissues.

Take Mascot peptide summary report of sample set 'dodo-repeat' as an illustration, considering PTMs of hydroxyproline, hydroxyglycine and deamidation, mass 1162.5906 Da was matched to at least five sequences as follows,

1) R.GQAGVMGFPGPK.G + Deamidation (NQ); Hydroxylation (P),

2) K.GDIGGPGFPGPK.G + 2 Hydroxylation (P); 2 Hydroxylation (G),

3) K.VDQVFGPRTK.C + Hydroxylation (P),

4) R.IGARMGRPEK.S + Hydroxylation (K); Hydroxylation (P); Hydroxylation (G),

5) K.ANNLFIVKSR.C + Deamidation (NQ)

Each of them contains different groups of PTMs, they greatly increase the difficulty of protein identification by database searching.

2.5.2 Detecting Hydroxylation in MS

Hydroxylation is one of the most common PTMs appear in collagen. As shown in Figure2.5.1, the process of proline hydroxylation adds 16 Da to the peptide mass.

Therefore the '16' gap could be regarded as symbol of hydroxylation. In MS spectrum, one hydroxylation corresponds to two peaks with 16 Da difference in mass, analogically, multiples of '16' gaps expose more than one hydroxylation. Take peak '1479.7877' in spots set 'dodo-repeat' as illustration. It is identified as collagen sequence 'GLHGEFGVPGPAGPR + 2 Hydroxylation (P)' (hydroxylation on the ninth and eleventh amino acids). In the same sample, peak '1463.7356' is identified as 'GLHGEFGVPGPAGPR + Hydroxylation (P)' (hydroxylation on the ninth amino acid), the same sequence as '1479.7877' but one less hydroxylation. These two peaks are 16 Da different in mass, where the '16 gap' appears.

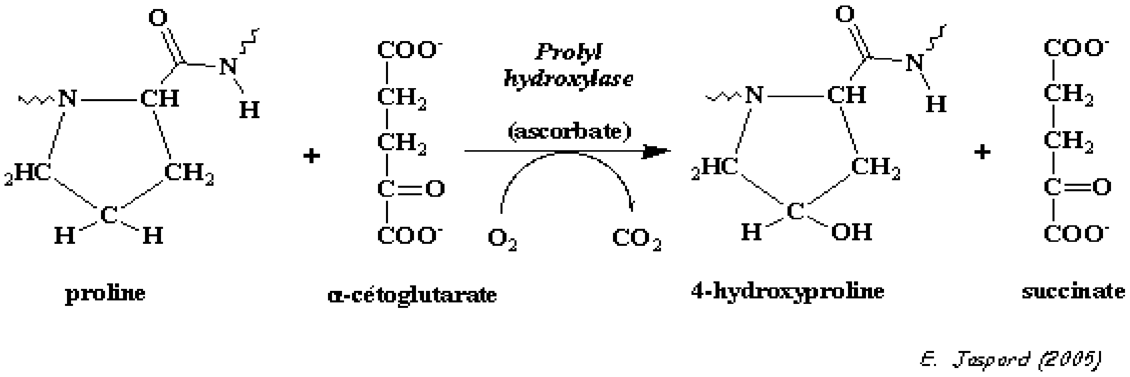


Figure 2.5.1 - Hydroxylation of proline

Based upon the '16' gap, a program was developed to detect peaks with one or multiple such gaps in MS spectrum. Suppose a peak in MS contains one hydroxylation, the peak on its left side with -16 Da difference would be very likely of the same sequence but without that hydroxylation. Therefore all peaks with multiple of 16 Da differences in mass are potentially correlated. By going through each peak in a MS spectrum, and searching for all peaks with 16, 32, 48, 64 and -16, -32, -48, -64 Da difference from it, a list of their intensity values (total number of observation) was worked out.

2.5.3 Hydroxylation Patterns

Using the above method, hydroxylation patterns can be obtained by plotting their intensity values. Taking peptide fragment 'AGAPGTPGPP' from chicken type I collagen for example, the peptide from two chicken samples (heated separately with 12℃ temperature difference) generated two similar peaks in MS (806.5 Da in 'AI' and 806.4 Da in 'AIV'). Their intensity of the '16' gaps are plotted in Figure 2.5.2 and Figure 2.5.3, which give similar patterns of hydroxyproline in these two samples. There are three Y-position amino acid prolines in their sequences, and the intensity of '3' is the highest, indicating those three prolines are all highly possible to be hydroxylated.

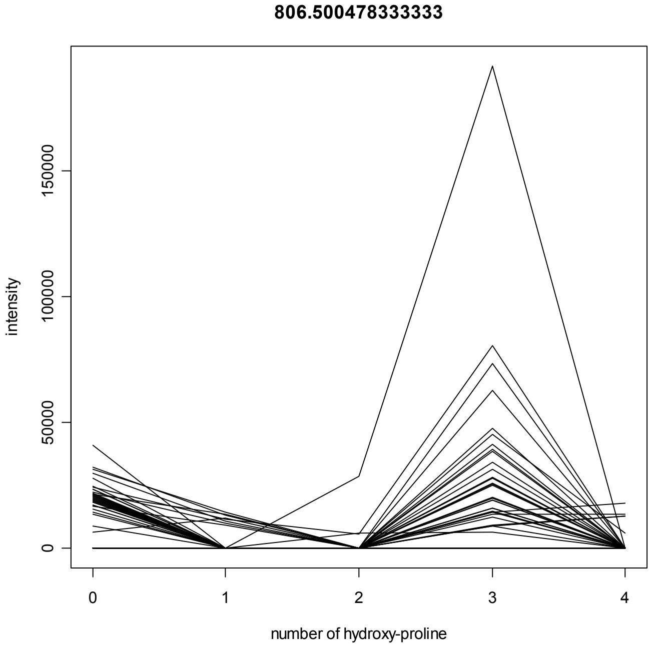


Figure 2.5.2 - Hydroxylation pattern for AI

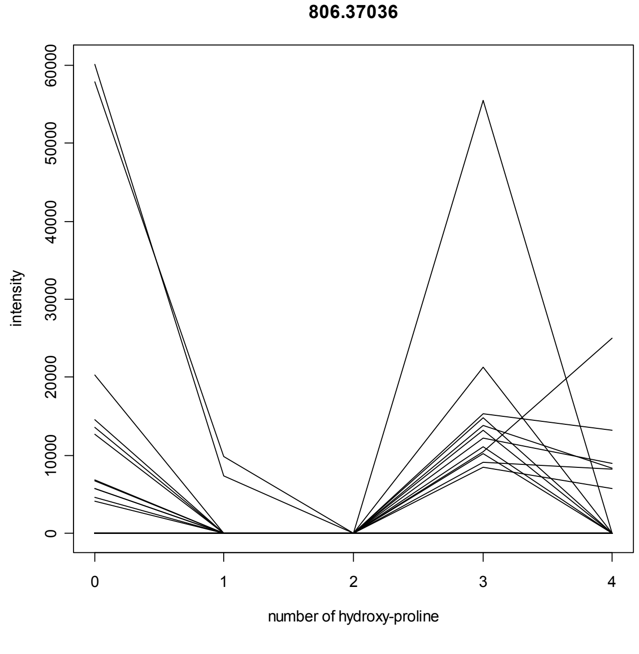


Figure 2.5.3 - Hydroxylation pattern for AIV

The two figures above show the hydroxylation intensity for same peptide from two chicken samples. The x-axis shows the number of '16' gaps included, where '1' to '4' corresponds to 16 to 64 Da difference to the target peptide, and '0' indicates the original peak; while y-axis shows intensity of peaks fit those gaps.

2.5.4 Internal Fragments and Hydroxylation Distribution

The relationship between internal fragments and hydroxylation intensity was investigated in this study. The ten sample sets in Table 2.4.2 were included, their hydroxylation intensity and the number of collagen markers for each peptide were plotted in Figure 2.5.4, where hydroxylation degree was found to increase with the number of collagen markers.

The distribution of collagen markers and hydroxylation demonstrated in the figure shows that, peptides with more collagen markers have higher hydroxylation degree. Accordingly, hydroxylation is more likely to take place in peptides with more collagen markers, suggested that collagen fragments are more possible to be hydroxylated because of abundance of proline and glycine.

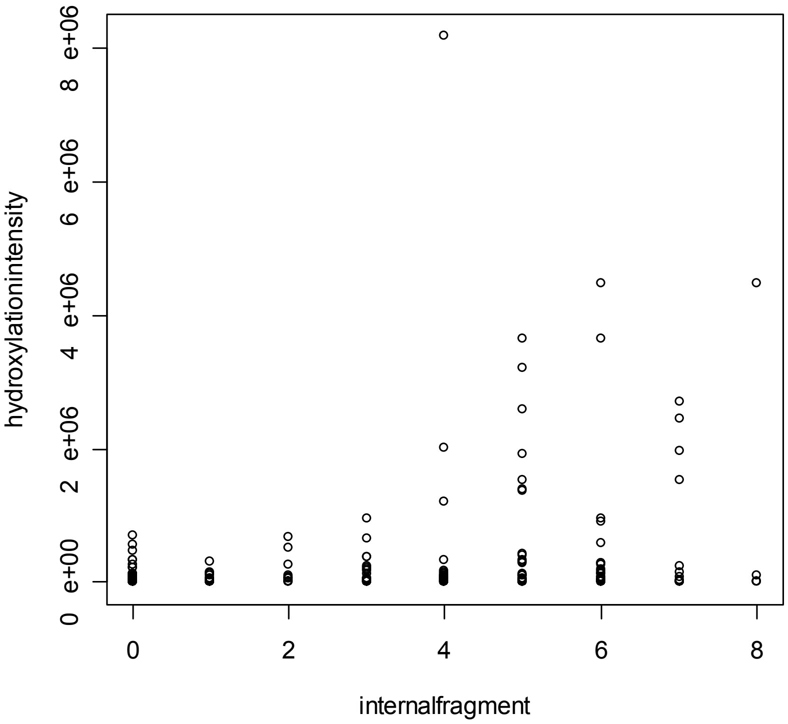


Figure 2.5.4- Distribution of collagen markers and hydroxylation

The figure shows the relationship between number of collagen markers and hydroxylation rate. The x-axis shows the number of internal collagen fragments, and y-axis gives the hydroxylation intensity detected for each peptide in the investigation.

Chapter 3

Database Building Methodology

The importance of a large database that contains sufficient numbers of collagen sequences has been highlighted in the preceding sections. Due to the deficiency of the existing database, this work aimed at establishing a theoretical type I collagen specific database named ‘UniColl’. This database was designed to include as many of hypothetical peptide sequences as possible for type I collagen.

3.1 Research Procedure

The mechanism of this database was to theoretically fragment tryptic peptide sequences from the data source and study the variations on these peptide fragments to generate a large database of theoretical peptide fragments which included all of the observed sequence variations found in the available sequences. The general procedure of building such a database can be demonstrated by a flow chart as in the Figure 3.1.1.

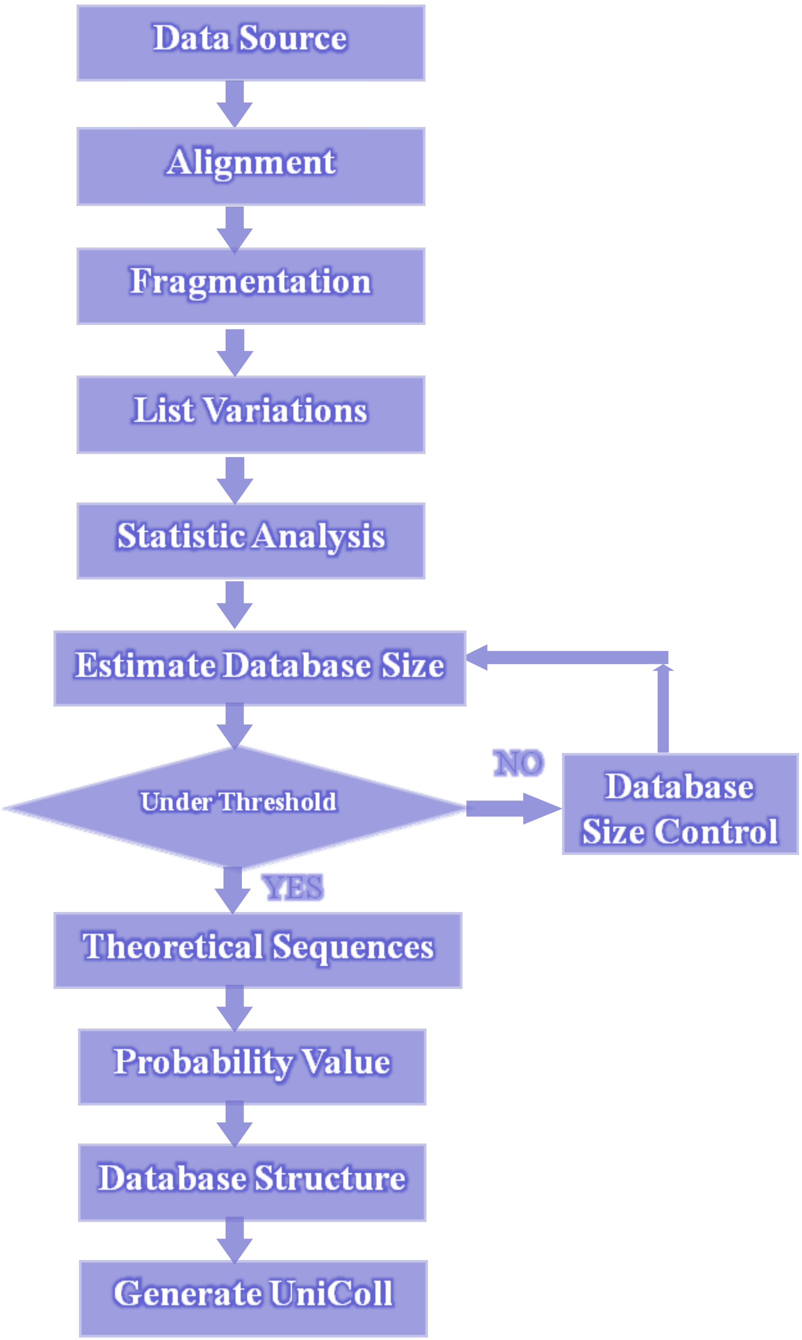


Figure 3.1.1 - The procedure of building UniColl

As demonstrated in the above figure, the first procedure of building UniColl is to collect type I collagen sequences from several available data sources including public and laboratory protein databases. These sequences would be selected as the data source for generating the database, and classified into the a1 and a2 chains, which are the two types of polypeptide chains composing type I collagen. By applying the multiple sequence alignment algorithm on each group, alignments of both a1 and a2 polypeptide chains can be produced. Tryptic fragments were then generated by making theoretical tryptic cleavage on both of the sequences. Taking each tryptic fragment as a unit, statistical analysis was applied on potential variations of its polypeptide sequence, with the varying points and presenting rate of each variation listed. The number of theoretical sequences generated by permutation and combination of these variations can then be estimated. If the number is in the reasonable scope, the group of theoretical sequences can be produced. Otherwise, if the number of theoretical sequences is out of the reasonable scope, it will be cut down to meet the limitation of the MASCOT search engine (which limits the maximum size of any protein). Once the number of these theoretical sequences is reduced to a reasonable size, these sequences will be concatenated and included as 'theoretical protein' in the database. An extra piece of information tagged to these sequences is a probability value of each theoretical sequence. This value can be calculated on the base of the presenting rate of each variation in that sequence, stands for its possibility of occurrence. All theoretical sequences grouped by tryptic fragmentation were then arranged in to the database, with each group ordering by their probability value from highest to lowest. As the result, data structure of the database was shaped and the database was established.

3.2 Methodology

3.2.1 Sequence Alignment

The molecule of most type I collagen consists of two α 1 and one α 2 polypeptide chains. Each of these polypeptide chains is composed of over one thousand amino acid residues, which are expressed as corresponding symbols arranging in a sequence. Such sequences for type I collagen collected from several data sources were initially classified into α 1(I) or α 2 (I) group, each of which need to be aligned as the foundation of sequence comparison among species. The basic procedure to compare a group of protein sequences is multiple sequence alignment, which aligns multiple sequences to minimise differences between sequences caused by amino acid substitutions.

The mechanism of multiple sequence alignment is the contrastive arrangement for the series of amino acid symbols according to their corresponding or substituting relationship. The purpose of such alignment is to locate the information in common from a group of homotypic proteins, in the form of a sequence similarity description that illustrates the conserved and varied parts in a group of sequences, in order to analyse their evolutionary or structural distinction.

The sequence alignment is the essential procedure for database building, as variations of amino acid substitutions in sequences can be observed and analyzed on this basis. While in the case of type I collagen, the alignment is more feasible due to the highly conserved composition of their sequences.

In this research, Geneious (version 4.7), an integrated bioinformatics software suite has been used to align the α 1(I) and α 2 (I) groups based on their unified GXY pattern. The aligning algorithm chosen in this work was MUSCLE, which includes fast distance estimation, progressive alignment and refinement (Edgar 2004).

3.2.2 Sequence Fragmentation

As the result of multiple sequence alignment, α 1(I) and α 2(I) alignments consist of 1000+ amino acid residues can be created. According to the mechanism of protein mass spectrometry analysis, which is also the basis of protein sequence database searching method, these two alignments need to be cleaved into shorter fragments following corresponding experimental principle.

In mass spectrometry, protein's polypeptide chains need to be cleaved into shorter peptide fragments by enzymolysis in order to fit the preferred mass range of instrumentation (typically < 5000 m/z) to produce information-rich spectra (Steen and Mann 2004). Therefore the protein MS outputs would be expressed as series of mass fingerprints from peptide fragments. Accommodating this feature, in the process of protein MS database searching, algorithms were designed to theoretically cleave protein sequences with various enzymes. On this basis, MS fingerprints for enzymatically-digested peptide fragments can be searched for the best-matched theoretical spectra in database generated using the selected enzyme.

In this piece of work, the endoprotease trypsin was selected. Trypsin is an aggressive and stable protease which is one of the most commonly used enzymes in protein mass spectrometry experiments. Trypsin cleaves proteins on the carboxy-terminal side of arginine and lysine residues unless followed by a proline. According to this digesting principle, theoretical tryptic cleavage can be applied on both α 1(I) and α 2(I) alignments, following the method of clipping after arginine (K) or lysine (R) residues with no proline (P) follows. As the result, either α 1(I) or α 2(I) alignment can be fragmentized into around 80 segments of 'tryptic' alignments, each of which contains under 40 amino acid residues. Compared to undigested original collagen polypeptide chain, these fragments will be more appropriate not only for the MS database searching, but also for generating theoretical sequences on the grounds of residue variations observed in each segment of alignment, because the number of possible combinations come from all variations present in an alignment would be more likely to be controlled in a reasonable scope from shorter sequences rather than longer.

3.2.3 Statistical Analysis on Variation

In order to produce abundant theoretical sequences to involve all amino acid residue variations present in the selected type I collagen sequences, statistical analysis would be applied on the alignments after theoretical tryptic fragmentation. Consider one segment of such fragmentized alignment as one unit, the elementary step of the statistical analysis is to record all residue variations observed in this unit. For this purpose, an m-by-n matrix 'A' was established for statistical data storage as follows.

$$\mathrm{A} = \begin{Bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{Bmatrix}$$

Formula 3.2.1

The two-dimensional array in this matrix consists of m rows and n columns, where 'n' represents the number of amino acid residues in the analyzed section, while 'm' represents the number of possible variations on each residue. Considering that residues would vary among the 20 types of amino acids which are the main composition of protein's polypeptide chain, the value for 'm' was assigned as 20 to include all possible variations. Elements of matrix A are denoted by the variable '$a_{i,j}$' which represents the element at row 'i' and column 'j' of the matrix, where the subscript 'i' varies from 1 to 20 and 'j' varies from 1 to n.

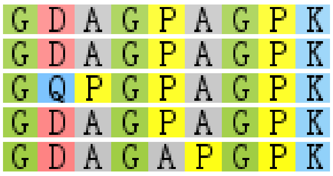
The value assignment for elements in matrix A would be used to record the amino acids attendance on corresponding residues. To realize such recording, value for '$a_{i,j}$' was assigned to register the appearing times of the No.i amino acid on the No.j residue. Here the correspondence of identifier 'i' and the 20 amino acid residues arranged in alphabetical order by their short codes is shown in Table. 3.2.1, while 'j'

simply refers to the specified residue's sort order in the analyzed piece of alignment.

Table 3.2.1 - List of amino acids

Identifier 'i'	Amino Acid	Short Code	Abbreviation
1	Alanine	A	Ala
2	Cysteine	C	Cys
3	Aspartic acid	D	Asp
4	Glutamic acid	E	Glu
5	Phenylalanine	F	Phe
6	Glycine	G	Gly
7	Histidine	H	His
8	Isoleucine	I	Ile
9	Lysine	K	Lys
10	Leucine	L	Leu
11	Methionine	M	Met
12	Asparagine	N	Asn
13	Proline	P	Pro
14	Glutamine	Q	Gln
15	Arginine	R	Arg
16	Serine	S	Ser
17	Threonine	T	Thr
18	Valine	V	Val
19	Tryptophan	W	Trp
20	Tyrosine	Y	Tyr

According to the procedure demonstrated above, a matrix can be set up for each piece of alignment for a tryptic peptide fragment cleaved from either α 1(I) or α 2(I) chain, with presenting rate recorded for every amino acid on each residue in the sequence. Take the following example to illustrate the particular process.



```
G D A G P A G P K
G D A G P A G P K
G Q P G P A G P K
G D A G P A G P K
G D A G A P G P K
```

Figure 3.2.1 - Example of a tryptic peptide alignment

The figure indicates the collagen sequence variation amongst five species in this tryptic peptide unit. The combinations made up from the variations will be demonstrated as follows.

As shown in Figure 3.2.1, assuming an alignment X' for a tryptic unit consists of 5 sequences each includes 9 amino acid residues. Set up a 20×9 matrix A' to record the statistical analysis data for this alignment as follows.

$$\mathrm{A'}=\left\{\begin{matrix} a'_{1,1} & a'_{1,2} & \cdots & a'_{1,9} \\ a'_{2,1} & a'_{2,2} & \cdots & a'_{2,9} \\ \cdots & \cdots & \cdots & \cdots \\ a'_{20,1} & a'_{20,2} & \cdots & a'_{20,9} \end{matrix}\right\} \qquad \textit{Formula 3.2.2}$$

By counting the appearing times of amino acid symbols in this piece of alignment, elements $a'_{i,j}$ of matrix A' can be assigned at the values shown in Table 3.2.2.

Table 3.2.2 shows the data storing structure of matrix A'. Cells $a'_{i,j}$ in the table record the appearing times of each amino acid residue through the alignment in Figure 3.2.1, while '0' means no show of certain amino acid on corresponding position in the alignment. For example, the assignment for cell $a'_{i,j}$ with $i = 3$ and $j = 2$ is valued at '4', that represents that aspartic acid (D) which is the No. 3 amino acid appointed at Table 3.2.1 appears 4 times on the second residue of the analysed alignment. Based on this form of integer matrix, theoretical sequences can be produced with their probability values calculated, which will be demonstrated in the following sections.

Table 3.2.2 - List of matix elements' value

$a_{i,j}$ i	j code	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$	$j = 6$	$j = 7$	$j = 8$	$j = 9$
$i = 1$	A	0	0	4	0	1	4	0	0	0
$i = 2$	C	0	0	0	0	0	0	0	0	0
$i = 3$	D	0	4	0	0	0	0	0	0	0
$i = 4$	E	0	0	0	0	0	0	0	0	0
$i = 5$	F	0	0	0	0	0	0	0	0	0
$i = 6$	G	5	0	0	5	0	0	5	0	0
$i = 7$	H	0	0	0	0	0	0	0	0	0
$i = 8$	I	0	0	0	0	0	0	0	0	0
$i = 9$	K	0	0	0	0	0	0	0	0	5
$i = 10$	L	0	0	0	0	0	0	0	0	0
$i = 11$	M	0	0	0	0	0	0	0	0	0
$i = 12$	N	0	0	0	0	0	0	0	0	0
$i = 13$	P	0	0	1	0	4	1	0	5	0
$i = 14$	Q	0	1	0	0	0	0	0	0	0
$i = 15$	R	0	0	0	0	0	0	0	0	0
$i = 16$	S	0	0	0	0	0	0	0	0	0
$i = 17$	T	0	0	0	0	0	0	0	0	0
$i = 18$	V	0	0	0	0	0	0	0	0	0
$i = 19$	W	0	0	0	0	0	0	0	0	0
$i = 20$	Y	0	0	0	0	0	0	0	0	0

Table 3.2.2 shows the data storing structure of matrix A' for the alignment in Figure 3.2.1, where ‘i’ is the index number of the amino acid in Table 3.2.1, while ‘j’ is the number of observation for the corresponding residue in the alignment. The grey cells indicate the nonzero values.

3.2.4 Generating Theoretical Sequences

For the purpose of enhancing the comprehensiveness for specific protein database, theoretical sequences would be created to involve as many variations as possible in type I collagen. The mechanism of generating such theoretical sequences is the combination of the observed amino acid substitutions, since they are very likely to occur in other unknown sequences but in different combinations. Take the alignment in Figure 3.2.1 for illustration, supposing the 5 sequences are all the information contained in database, while there is a peptide with sequence of 'GDPGPAGPK', with popular variations of 'D' and 'P' for the second and third residues but the combination has not appeared in the 5 sequences in database. In order to get the correct sequence matching in database for this peptide, the most efficient way is to include this sequence in database. This is applicable on the basis of the statistical analysis applied on the alignment that is recorded in relevant matrix A' illustrated above. The methodology used for generating such theoretical sequences to replenish the database is the multiplication principle.

The multiplication principle, also know as the rule of product, is a fundamental counting principle in combinatorics. The idea of this principle is that if there are n steps of doing something, while there are m_1 ways of doing step 1, m_2 ways of doing step 2, ... , m_n ways of doing step n, then there are $m_1 \times m_2 \times \cdots \times m_n$ ways of performing all actions.

As applying this principle on protein sequences, considering a piece of peptide fragment X consists n amino acid residues in a sequence of $x_1 x_2 ... x_n$, there are n steps to obtain the sequence of X, that is to determine the amino acid on $x_1, x_2, ..., x_n$ one by one. Assuming there are m_1 options for x_1, m_2 options for x_2, ... , m_n options for x_n, using C_X to represent the number of possible combinations to form the sequence of X, then

$$C_X = m_1 \times m_2 \times \cdot \times m_n \qquad \textit{Formula 3.2.3}$$

In the case of the fragmentized type I collagen peptide alignments involved in this research, in order to generate theoretical sequences to include all variations, each amino acid appearing on a position in the alignment would be considered as one option for that position. Take the 5 sequences alignment X' in Figure 3.2.1 as an example, set the 9 residues as $x'_1, x'_2, ..., x'_n$. Options for each residue in this alignment have been recorded in matrix A', while number of options for $x'_j\ (j = 1,2,...,9)$ is equivalent to the number of nonzero elements in the column vector $A'[j]$ as follows:

$$\mathrm{A}'[j] = \left\{ \begin{matrix} a'_{1,j} \\ a'_{2,j} \\ \vdots \\ a'_{20,j} \end{matrix} \right\} \qquad \textit{Formula 3.2.4}$$

In order to count the number of nonzero elements in $A'[j]$, taking an integer variable 'b_j' as an enumerator, a traversal would be applied on elements of $A'[j]$. Once an element being visited in the traversal is nonzero, take the procedure of b_j=b_j+1, until all elements have been visited. Therefore, the number of options m'_j for residue $x'_j\ (j = 1,2,...,9)$ can be expressed as follows:

$$m'_j = b_j (j = 1,2,\cdots 9) \qquad \textit{Formula 3.2.5}$$

To illustrate the specific procedure of statistical analysis of the alignment X', variations observed on residues $x'_j\ (j = 1,2,...,9)$ can be demonstrated in Table 3.2.3.

Table 3.2.3 - Variations in the alignment

x'_j : residue orders	x'_1	x'_2	x'_3	x'_4	x'_5	x'_6	$x'_{'7}$	x'_8	x'_9
variations on x'_j	G	D,Q	A,P	G	A,P	A,P	G	P	K
m'_j : number of options	1	2	2	1	2	2	1	1	1

The first row shows the orders for the nine residues; the second row shows the observed variations on each residue; the third row shows the number of variations.

As in the above table, each amino acid present at the residue $x'_j\ (j = 1,2,...,9)$ can be considered as one option for the corresponding x'_j. For example A and P are present at x'_3, so that there are 2 options for x'_3. According to Formula 3.2.3, set $C_{X'}$ to represent the number of theoretical sequences for the alignment X', $C_{X'}$ can be calculated as follows:

$$C_{X'} = m'_1 \times m'_2 \times \cdots \times m'_9 = 1 \times 2 \times 2 \times 1 \times 2 \times 2 \times 1 \times 1 \times 1 = 16 \qquad \textit{Formula 3.2.6}$$

Therefore, 16 theoretical sequences can be generated from combinations of residue variations observed in alignment X' and listed as follows:

GDAGAAGPK; GDAGAPGPK; GDAGPAGPK; GDAGPPGPK
GDPGAAGPK; GDPGAPGPK; GDPGPAGPK; GDPGPPGPK
GQAGAAGPK; GQAGAPGPK; GQAGPAGPK; GQAGPPGPK
GQPGAAGPK; GQPGAPGPK; GQPGPAGPK; GQPGPPGPK

The database of 'UniColl' is designed to contain sequences such as the above 16 ones, which including the original 5 sequences which were used as the data source, while generating 11 new theoretical sequences to cover other possible combinations come from the present variations which might occur in an unknown peptide sequence. For example, the peptide with sequence of 'GDPGPAGPK' mentioned at the beginning of this section then would be more likely to be identified in 'UniColl', while other protein databases cannot export a correct searching result because that sequence is not included.

3.2.5 Probability Value

For the purpose of arranging the enormous number of theoretical sequences, certain ordering rules need to be developed.

In conventional protein databases such as ‘UuiProt’, protein sequences are normally recoded as series of symbols, standing for amino acid residues aligning in the polypeptide chain of a certain type of protein from a particular species, while each symbol is assigned with a serial number which denotes the position of the corresponding residue. For instance, if the database searching outcome includes sequence ‘ProteinOne_0010-0020’, that means the peptide located from the 10th residue to the 20th residue of the ProteinOne’s sequence has been matched to the sample.

However in the database of UniColl, groups of theoretical sequences were generated on the basis of the tryptic peptide units. These mathematically combined sequences are neither necessarily corresponding to a specific species, nor even existing in the nature. Therefore it is meaningless to assign them to certain species. Considering this situation, the most feasible way of arranging the theoretical sequences in UniColl would be grouping by tryptic peptide units, which is to classify sequences from the same of peptide unit into the same ‘protein’ group. By alining sequences in each group into a longer synthetic peptide (treated by MASCOT as a ‘protein’), every peptide unit generated a sequence containing all theoretical combinations generated from it. Based on the database-searching algorithm, such a long sequence would be cleaved into tryptic peptide fragments and matched to the input peptide sample separately; therefore the ‘UniColl protein’ would be treated as a normal protein, with the most likely peptide being surfaced.

Considering that the number of theoretical sequences in each peptide unit was still large, the ordering principle for sequences inside a group was needed to be figured out as well. The output format of the MASCOT database searching that shows the serial numbers of the matched peptide indicated the location of that peptide (such as ‘ProteinOne_0010-0020’). In view of sequences in UniColl were not aligned in the order of nature protein chains, but series of theoretic sequences from same peptide unit, the ordering rule should ideally be designed to make the displayed serial numbers meaningful other than only telling the location.

Since sequences in UniColl were generated from theoretically combination of all possibilities exist in data source, some of the sequences commonly exist in several species, conversely most will be absent in nature. In the case of database searching, if the sample is matched to two sequences with the equal score, the occurring probability could be considered as a referential score help with deciding which match is more likely to be the actual sequence of the sample. However if the theoretical sequences with different occurring probabilities are arranged randomly, it will be difficult to tell the rare ones from the common ones. Therefore, a group of theoretical sequences could be arranged into a long sequence ordered by their probability values from high to low. In this case, the serial number of each peptide shown in the database searching result would be correlated to its occurring probability, while the smaller serial numbers signify the higher probability value of the relevant sequence occurring in type I collagen samples.

The probability value of each theoretical tryptic peptide sequence could be calculated as the product of the probability value of each amino acid in the sequence, while the probability value of an amino acid here could be computed as the frequency of its existence at that position through all species in the alignment. The calculation could be practised on the basis of statistical analysis of residues variation. Assuming in a theoretical sequence T which is composed of n residues, amino acid variation V_i ($i = 1, 2, ..., n$) turns up P_i times on the ith residue through the source alignment which totally contains m species, then the probability value P of sequence T can be computed as

$$P = \prod_{i=1}^{n} \frac{P_i}{m} \qquad \textit{Formula 3.2.7}$$

Still take the 5 sequences alignment in Figure 3.2.1 as an example, where $n = 9$ and $m = 5$. The presenting times P_i for the theoretical sequence 'GDAGAAGPK' generated from this alignment can then be calculated through simply statistical analysis as $P_i = [5,4,4,5,1,4,5,5,5]$, thereafter the probability value P for this sequence can be computed as follows.

$$P = \prod_{i=1}^{9} \frac{P_i}{5} = 1 \times \frac{4}{5} \times \frac{4}{5} \times 1 \times \frac{1}{5} \times \frac{4}{5} \times 1 \times 1 \times 1 = 0.1024 \qquad \textit{Formula 3.2.8}$$

Further more, the probability value for all the 16 theoretical sequences from this peptide fragment can be computed as in Table 3.2.4.

Table 3.2.4 - Probability value for theoretical sequences

Theoretical Sequences	P_i									P
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$	
GDAGAAGPK	5	4	4	5	1	4	5	5	5	0.1024
GDAGAPGPK	5	4	4	5	1	1	5	5	5	0.0256
GDAGPAGPK	5	4	4	5	4	4	5	5	5	0.4096
GDAGPPGPK	5	4	4	5	4	1	5	5	5	0.1024
GDPGAAGPK	5	4	1	5	1	4	5	5	5	0.0256
GDPGAPGPK	5	4	1	5	1	1	5	5	5	0.0064
GDPGPAGPK	5	4	1	5	4	4	5	5	5	0.1024
GDPGPPGPK	5	4	1	5	4	1	5	5	5	0.0256
GQAGAAGPK	5	1	4	5	1	4	5	5	5	0.0256
GQAGAPGPK	5	1	4	5	1	1	5	5	5	0.0064
GQAGPAGPK	5	1	4	5	4	4	5	5	5	0.1024
GQAGPPGPK	5	1	4	5	4	1	5	5	5	0.0256
GQPGAAGPK	5	1	1	5	1	4	5	5	5	0.0064
GQPGAPGPK	5	1	1	5	1	1	5	5	5	0.0016
GQPGPAGPK	5	1	1	5	4	4	5	5	5	0.0256
GQPGPPGPK	5	1	1	5	4	1	5	5	5	0.0064
SUM	--									1

In the above table, column2 to column10 express occurring times P_i on the *i*th residue of the corresponding theoretical sequence in column1, while the last column shows the computation results of the probability value *P*. As shown in the last row, the sum of probability values becomes 1, confirming the mathematical principle that the total probability of all possibilities of an event should be 1.

By reordering the 16 sequences by their probability values, a ranking comes up as in Table 3.2.5.

Table 3.2.5 - Reordering the theoretical sequences by P value

Rank	Theoretical Sequences	*P*
1	GDAGPAGPK	0. 4096
2	GDAGAAGPK	0. 1024
3	GQAGPAGPK	0. 1024
4	GDPGPAGPK	0. 1024
5	GDAGPPGPK	0. 1024
6	GQPGPAGPK	0. 0256
7	GQAGPPGPK	0. 0256
8	GQAGAAGPK	0. 0256
9	GDPGPPGPK	0. 0256
10	GDPGAAGPK	0. 0256
11	GDAGAPGPK	0. 0256
12	GQPGPPGPK	0. 0064
13	GQPGAAGPK	0. 0064
14	GQAGAPGPK	0. 0064
15	GDPGAPGPK	0. 0064
16	GQPGAPGPK	0. 0016

In the above table, the first column is the ordering number; the second column shows the theoretical sequence combinations; while the third column gives the *P* value of each combination.

According to the probability values, higher rank in the above Table suggests higher possibility that theoretical sequence presenting in the relevant protein. In this case, the rank 1 sequence ‘GDAGPAGPK’ is obviously more popular than the others. Coming back to the original data source, the alignment in Figure 3.2.1, sequence ‘GDAGPAGPK’ is truly the most popular one by occurring 3 times out of the 5 sequences. While the rank 16 sequence ‘GQPGAPGPK’ presents a rare combination which is very unlikely to occur including in the 5 data sources. On this basis, once the database searching matched the sample to the rank 1 and rank 16 sequences at the same time both with the highest score, then the P value would suggest that the rank 1 sequence is more likely to be the correct match.

Concretely, as the 16 sequences could be alined to a synthetic peptide sequence according to the order of value P from high to low, which give out a long sequence as ‘GDAGPAGPKGDAGAAGPKGQAGPAGPKGDPGPAGPKGDAGPPGPKGQPGP AGPKGQAGPPGPKGQAGAAGPKGDPGPPGPKGDPGAAGPKGDAGAPGPKG QPGPPGPKGQPGAAGPKGQAGAPGPKGDPGAPGPKGQPGAPGPK’.

In database searching, these 144 residues in this sequence will be assigned serial numbers from 1 to 144. And the serial number would be shown in the searching result, such as the ‘Peptide_001-009’ stands for the rank 1 sequence ‘GDAGPAGPK’, while the ‘Peptide_136-144’ stands for the the rank 16 sequence ‘GQPGAPGPK’. If ‘Peptide_001-009’ and ‘Peptide_136-144’ both appeared in the database searching outcome list with the highest score, then the former one would be suggested to be a better match rather than the latter one.

However, the P value should be only considered as a referential parameter, not the absolute judging standard. Besides, the P value is not necessarily equals to the actual probability of the relevant sequence occurs. For example the P value of the rank 1 sequence ‘GDAGPAGPK’ is 41%, but the real probability that this sequence occurred in the 5 data sources is 60%. Therefore, the ranking and comparison of P values among all theoretical sequences in an integrated alignment has more significance than the absolute P value itself.

3.2.6 Data Storage

The data storage of protein databases is basically in the form of sequence of symbols that represent protein's amino acid composition. As metioned in the preceding section, conventional protein database stores such sequences in the unit of protein molecule, which means one sequence corresponds to one protein molecule. However in the 'UniColl' database, sequences are stored in the unit of tryptic peptides, which means one sequence is a combination of the theoretical sequences generated from the same piece of tryptic peptide unit. This model of data storage is applicable for database searching under the mode of tryptic enzymolysis, since sequences would be cleaved into tryptic fragments before matching to the input data, therefore a synthetic sequence for peptides represents the tryptic cleaved products in database searching process.

Besides, considering the significance of the probability value discuss above, theoretical peptide sequences aligning in one synthetic sequence should be ordered by their P value from high to low. Such synthetic theoretical sequences were named by the location of their source peptide fragment in the relevant polypeptide chain, for the purpose of distinguishing theoretical sequences come from different sections of the collagen molecule. The sequences were written into text files in FASTA file format which can be interpreted by the MASCOT searching engine, and these made up the UniColl database.

Take the alignment X' from Figure 3.2.1 for example. Assuming the alignment covers from the 1st amino acid residue to the 9th in the α 1(I) chain, then the synthetic sequence conjoining the 16 theoretical sequences demonstrated at the end of *Section 3.2.5* would be named as 'α 1(I) 0001-0009', then the rank 1 peptide ‘GDAGPAGPK’ will be 'α 1(I) 0001-0009_Peptide_001-009’.

3.2.7 Database Size Control

Keeping database size under control is necessary in database construction and maintenance. In computing of theoretical combinations, several peptide units produce too large amount of sequences to be loaded by UniColl. In order to control the size of UniColl to a manageable scale, a threshold $'T'$ need to be set as a reasonable scope for the data capacity. If the number of theoretical sequences generated from an alignment section exceeds the threshold $'T'$, certain measures would be adopted to cut down the number. Considering that different tryptic peptides (the basic units of the database searching procedure) consist different number of amino acid residues, the threshold would be set specific to the number of theoretical peptide fragments rather than the number of residues for a whole conjoined sequence. This principle can be demonstrated in the following formula, where C_X represents the number of possible combinations to form the sequence of alignment X as in Formula 3.2.3.

$$C_X \leq T \qquad \textit{Formula 3.2.9}$$

In this piece of work, the measure to minimize the database size when it achieves this threshold is to ignore the least popular variations, and the measure would be repeated until reducing the database size under threshold. This is in consideration of that the combination with uncommon variations would be less likely to occur in an unknown collagen sequence compared to those contain more common variants, or indeed the uncommon variants might come from wrong sequencing or alignment. The method of getting the least popular variation is to compare the sum of appearing times recorded in row vectors of matrix A. Set the S_i as the sum of the No. i amino acid's attendance times in a tryptic fragmentized alignment with n residues, then the No. i amino acid with the corresponding S_i of the minimum nonzero value among all residues is the least popular variation in this alignment. As shown in Formula 3.2.10, value of S_i can be simply calculated by sum up element $a_{i,j}$ which records the appearing times of relevant amino acid.

$$S_i = a_{i,1} + a_{i,2} + \cdots + a_{i,n} \qquad \textit{Formula 3.2.10}$$

Still take the alignment X' from Figure 3.2.1 for example, assuming the threshold set for database is $T'=10$. The number of theoretical sequence generated from alignment X' was calculated in Formula 3.2.3 as $C_{X'}=16$. In order to meet the condition in Formula 3.2.9, database size control measure would be executed with the first step of searching for the least popular variation. The minimum nonzero $S_i (i = 1,2,\cdots,9)$ is found to be $S_{14}=1$, therefore variation 'Q' which existed only once in this alignment would be ignored.

Theoretical combinations generated from alignment X' without including 'Q' as a variation will then be reduced to 8 sequences that are listed as follows:

GDAGAAGPK; GDAGAPGPK; GDAGPAGPK; GDAGPPGPK

GDPGAAGPK; GDPGAPGPK; GDPGPAGPK; GDPGPPGPK

In this situation, $C_{X'}=8$ and $T'=10$, therefore the condition of $C_X \leq T$ is satisfied, and no more database size reducing measures need be repeated.

Chapter 4

Building UniColl

The procedure of building the UniColl database has been illustrated as in Figure 3.1.1 and methodologies have been explained in the rest sections of Chapter 3. The specific progress of building the UniColl database is demonstrated in this chapter.

4.1 Data Source

The UniColl database was originated from the alignment of a group of type I collagen sequences. In order to embrace as comprehensive data source as possible according to the existing information, data mining has been applied on the main public proteomic databases including UniProt and NCBI, as well as several protein databases from academic laboratories such as BioArCh and UCSC. These main data sources are introduced below.

The universal protein database 'UniProt' is one of the main public species-specific resources of protein data created by Swiss-Prot, TrEMBL and PIR.. The National Center for Biotechnology Information (NCBI) advances science and health by providing access to biomedical and genomic information. The 'UCSC' refers to the University of California Santa Cruz which organized a gene bank including numbers of type I collagen sequences. And the BioArCh Laboratory of York University sequenced a variety of type I collagen samples by mass spectrometry integrated with EST technique.

As the outcome of data source searching, 40 species with their type I collagen sequences highly covered in either published or unpublished databases were sorted out as the original data for building UniColl. The sequences consist of 36 α1 chains and 38 α2 chains that cover 40 vertebrate species including 24 mammals, 3 birds, 2 amphibians, 2 reptiles and 9 fishes. The reason not including invertebrate species is that 1) few type I collagen sequence information covered in current database for invertebrate species; 2) their sequences are more variable compared with vertebrate species; 3) the research target of this work in focus on the fossil collagen, which mostly come from vertebrate species.

Detailed information of the collected sequences is illustrated in Table.4.1.1, which lists names and sources for the 74 type I collagen sequences of α1(I) and α2(I) chains from the 40 species which were involved in UniColl data source.

Table 4.1.1 - List of data source

Species	α1(I)	Source	Reference	α2(I)	Source	Reference
Human	√	Published	P02452	√	Published	P08123
Chimp	√	UCSC GB	panTro2	√	UCSC GB	panTro2
Rhesus	√	Published	gi|109114305	√	Published	gi|109104853
Galago	√	UCSC GB	otoGar1	√	UCSC GB	otoGar1
TreeShrew	√	UCSC GB	tupBel1	√	UCSC GB	tupBel1
Mouse	√	Published	P11087	√	Published	Q01149
Rat	√	Published	A3KNA1	√	Published	P02466
Guineapig	√	UCSC GB	cavPor2	√	UCSC GB	cavPor2
Rabbit	√	UCSC GB	oryCun1	√	UCSC GB	oryCun1
Dog	√	Published	Q9XSJ7	√	Published	gi|50978939
Cat	√	UCSC GB	felCat3	√	UCSC GB	felCat3
Horse	√	UCSC GB	equCab1	√	UCSC GB	equCab1
Sheep	√	BioArCh	Sheep	√	BioArCh	Sheep
Goat	X	N/A	N/A	√	BioArCh	Goat

Species	α1(I)	Source	Reference	α2(I)	Source	Reference
Pig	√	BioArCh	Pig	√	BioArCh	Pig
Cow	√	Published	P02453	√	Published	P02465
Shrew	√	UCSC GB	sorAra1	√	UCSC GB	sorAra1
Hedgehog	√	UCSC GB	eriEur1	√	UCSC GB	eriEur1
Armadillo	√	UCSC GB	dasNov1	√	UCSC GB	dasNov1
Tenrec	√	UCSC GB	echTel1	√	UCSC GB	echTel1
Elephant	√	UCSC GB	loxAfr1	√	UCSC GB	loxAfr1
Mammuthus	√	BioArCh	Mannuthus	√	BioArCh	Mannuthus
Opossum	√	UCSC GB	monDom4	X	N/A	N/A
Platypus	√	UCSC GB	ornAna1	√	UCSC GB	ornAna1
Ostrich	X	N/A	N/A	√	BioArCh	Ostrich
Chicken	√	Published	P02457	√	UCSC GB	galGal3
Dodo	√	BioArCh	Dodo	√	BioArCh	Dodo
Giant tortoise	√	BioArCh	GT	√	BioArCh	GT
Green Anole	√	UCSC GB	anoCar1	√	BioArCh	anoCar1
Xenopus	√	Published	gi|148222553	√	Published	gi|118404410
Frog	√	Published	gi|3242649	√	Published	O93484
Fugu	√	UCSC GB	fr2	√	UCSC GB	fr2
Tetraodon	√	UCSC GB	tetNig1	√	UCSC GB	tetNig1
Stickleback	√	UCSC GB	gasAcu1	√	UCSC GB	gasAcu1
Medaka	√	UCSC GB	oryLat1	√	UCSC GB	oryLat1
Halibut	√	Published	Q5NT96	√	Published	gi|56565283
Zebrafish	√	UCSC GB	danRer4	√	Published	gi|47937807
Ray Raja kenojei	√	Published	Q4W6W6	X	N/A	N/A
Trout	X	N/A	N/A	√	Published	gi|14164349
Keta	X	N/A	N/A	√	Published	AB075699

In the above table, the "source" column, "Published" indicates to sequences which are published in public databases, “UCSC GB” means unpublished sequences from UCSC (University of California Santa Cruz) gene bank, while “MS+EST” refers to sequences generated from York University BioArCh Laboratory by mass spectrometry sequencing integrated with EST technique. In the "α1(I)" and "α2(I)" columns, tick “√” means the sequence is covered, while cross “X” means the sequence is not available yet. The reference is the name of sequence in database, and “N/A” means the absence of relevant sequences.

During the analysis of collagen sequences data source, it was found that several sequences contain very limited information, most of which was repetition from other fully covered sequences. The probable reason for this was because the sequences came from matches obtained from MS database searching, consequently they are identical to the source sequence in database they were matched to. Considering these sequences contained too little information, which is additionally not exclusive to reveal unique variations, therefore these low quality sequences were removed from the data source for theoretical sequence generating. After that, 28 sequences with high capacity of information were reserved for the following analysis.

4.2 Sequence Alignment

As the outcome of data source searching, 74 pieces of type I collagen sequences from 40 species were collected. Although the data source covered very limited sequences, fortunately their arrangements were highly conserved due to the consistent G-X-Y molecular structure of type I collagen. Intriguingly the position of tryptic cleavage sites, arginine (R) and lysine (K) are remarkably consistent through most species especially for mammals; this provides a convenient set of references for their alignment.

Using the software tool Geneious version 4.7 to approach the alignment, the 36 α1 chains for type I collagen can be arranged into an alignment composed of 1057 amino acid residues while the alignment of α2 chains contained 1041 residues, which excluded the propeptides and telopeptides on each terminal. The alignments were checked over under the principle of ensuring the consistency of arginine (R) and lysine (K). The aligning results are presented in Figure 4.2.1 to Figure 4.2.10, each of which consecutively displays a 200 residues section out of the alignment. The original alignments were recorded in Appendix 1 and Appendix 2, which were converted from FASTA files to facilitate reading.

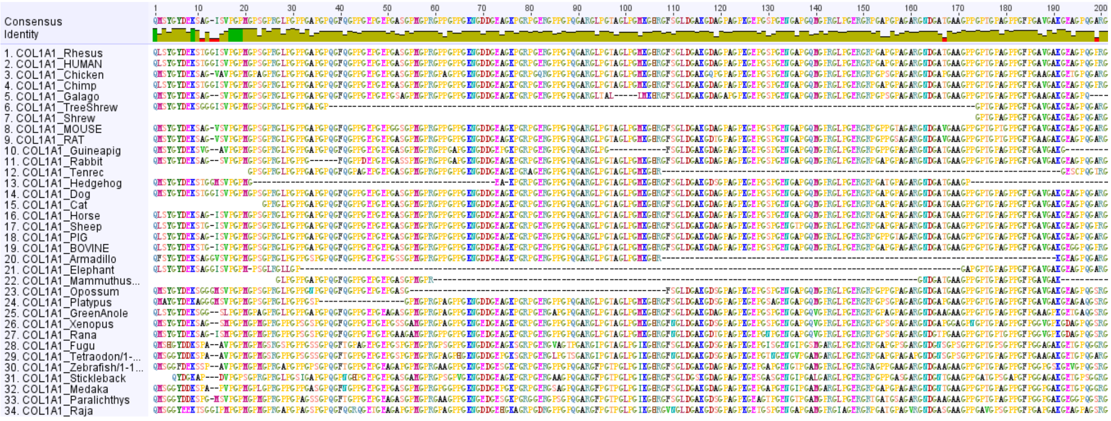


Figure 4.2.1 - Alignment of type I collagen α1 chains 0-200

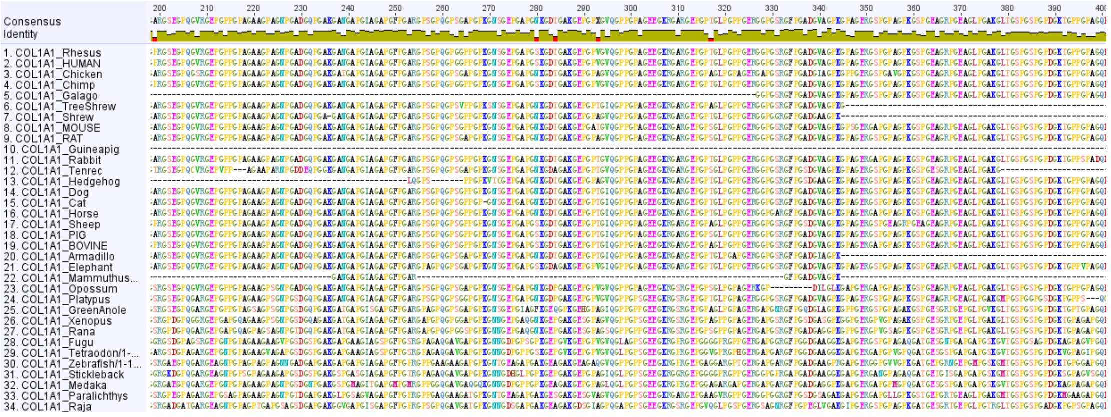


Figure 4.2.2 - Alignment of type I collagen α1 chains 200-400

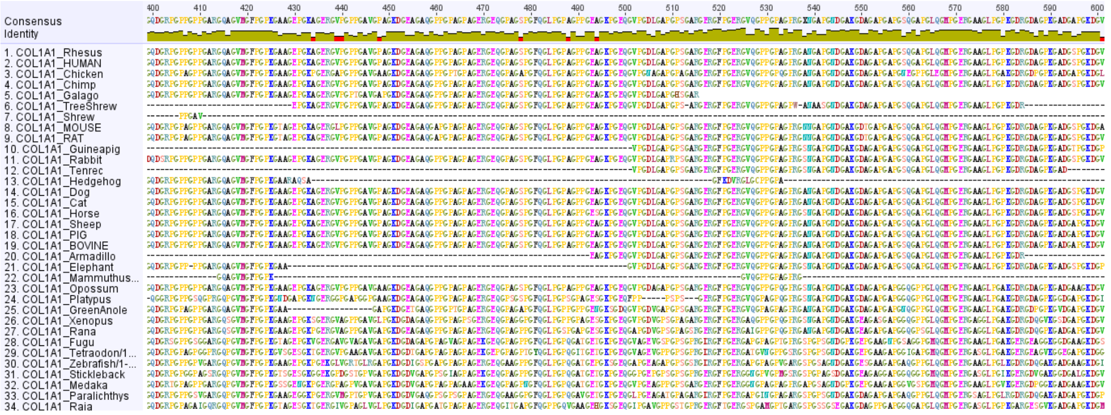


Figure 4.2.3 - Alignment of type I collagen α1 chains 400-600

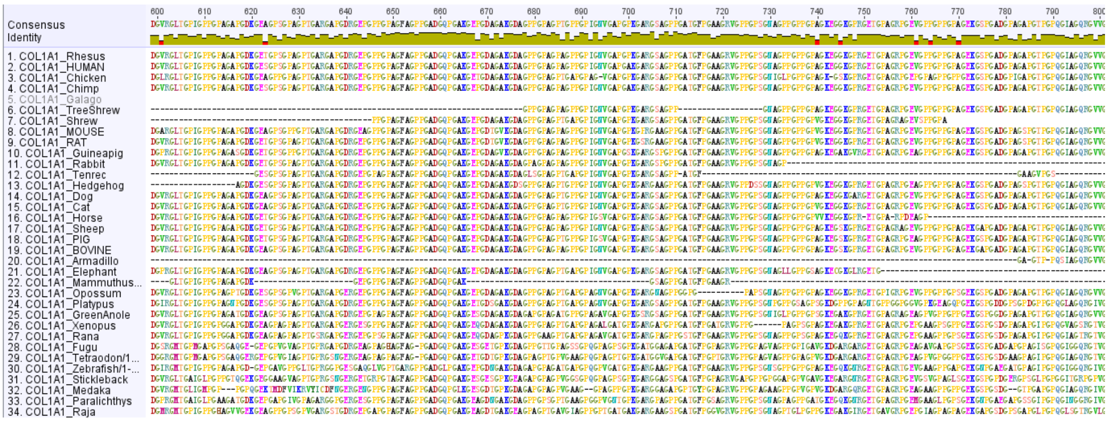


Figure 4.2.4 - Alignment of type I collagen α1 chains 600-800

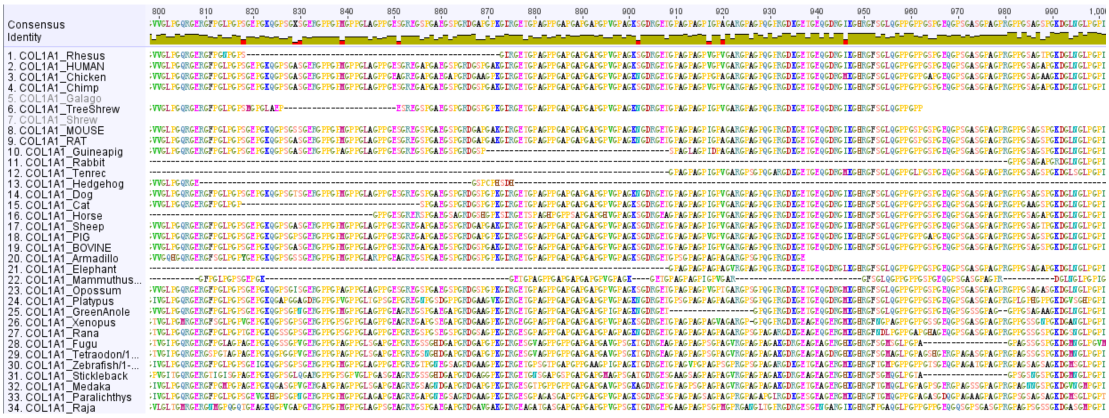


Figure 4.2.5 - Alignment of type I collagen α1 chains after 800

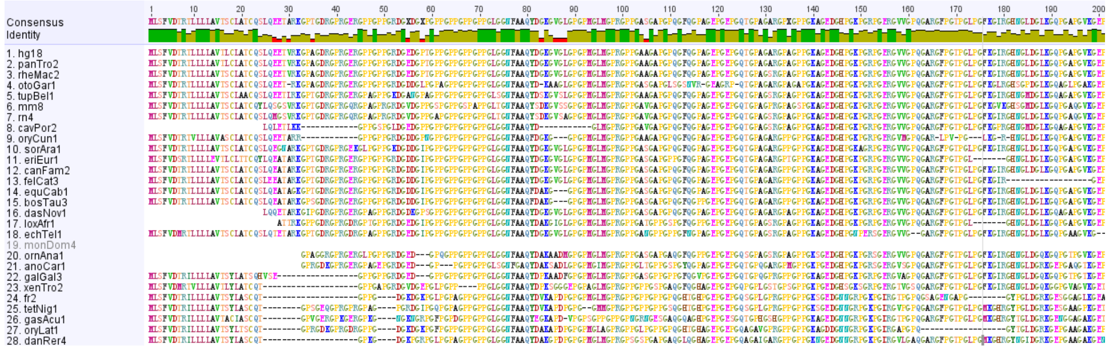


Figure 4.2 6 - Alignment of type I collagen α2 chains 0-200

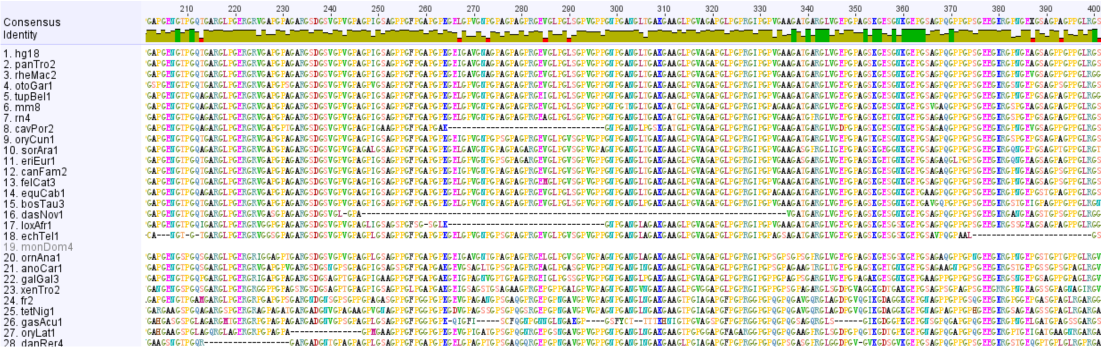


Figure 4.2.7 - Alignment of type I collagen α2 chains 200-400

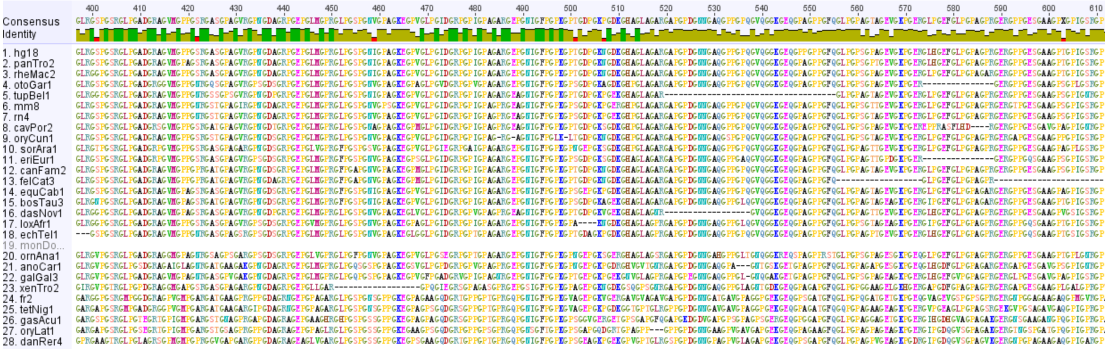


Figure 4.2.8 - Alignment of type I collagen α2 chains 400-600

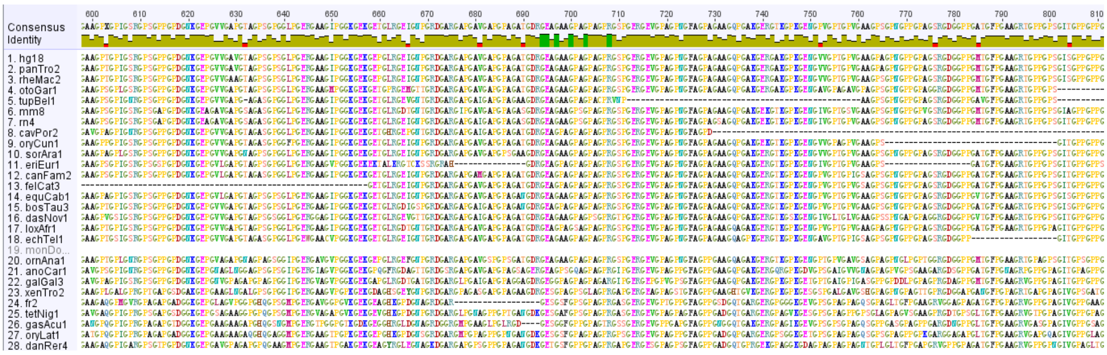


Figure 4.2.9 - Alignment of type I collagen α2 chains 600-800

Figure 4.2.10 - Alignment of type I collagen α2 chains after 800

Figures 4.2.1 to 4.2.10 displayed the Geneious alignments in the unit of 200 amino acid residues for type I collagen sequences from the species in Table 4.1.1. Figure 4.2.1 to 4.2.5 show the alignment of α1 chains with 1057 residues, and Figure 4.2.6 to 4.2.10 show the alignment of α2 chains with 1041 residues. The original alignments are readable in Appendix 1 and Appendix 2, and these figures are displayed here for general view of the alignment patterns only.

4.3 Sequence Fragmentation

According to the sequence fragmentation algorithm explained in *Section 2.2.3*, α 1(I) and α 2(I) sequence alignments consist of 1000+ amino acid residues would be cleaved into shorter peptide fragments. Applying virtual tryptic digestion to the alignments, α 1(I) or α 2(I) chains are cleaved into peptide fragments with their sequences ending up with arginine (K) or lysine (R). Considering the preferred mass range of mass spectrometry input, peptide that consists less than six amino acids would be less likely identified from the sample (and contain the least information), therefore such small piece of peptide were not included in generating the theoretical sequences for UniColl.

As the result, the alignment for α1 chain could be cleaved into 91 tryptic peptide fragments, 70 of which are not shorter than 6; while the α2 alignment created 86 such units, in which 65 ones have over 6 amino acids. The longest units include 38 residues in both α1 and α2 chains. The results of fragmentation are shown in Table 4.3.1 and Table 4.3.2 with all tryptic units listed for both the α1(I) and α2(I) chains. Further information is available in Appendix 3.

Table 4.3.1 - Fragmentation for a1(I)

Serial No.	N-Term	C-Term	End With	Length	Variations
1	1	9	K	9	1.08E+02
2	10	26	R	17	3.89E+04
3	27	59	R or Q	33	8.49E+07
4	60	67	K or E	8	2.88E+02
5	68	76	K	9	3.60E+01
6	77	79	R	3	--
7	80	83	R	4	--
8	84	92	R	9	6.00E+01
9	93	104	K	12	1.20E+01
10	105	107	R	3	--
11	108	116	K	9	4.00E+00
12	117	125	K	9	3.00E+00
13	126	143	R	18	2.30E+03
14	144	149	R	6	1.00E+00
15	150	151	R	2	--
16	152	161	K or R	10	1.92E+02
17	162	191	K	30	6.64E+06
18	192	200	R	9	1.28E+02
19	201	209	R	9	6.40E+01
20	210	236	K	27	4.15E+04
21	237	254	R	18	2.16E+03
22	255	269	K	15	6.91E+03
23	270	281	K	12	2.88E+03
24	282	287	K	6	2.40E+01
25	288	307	K	20	1.15E+03
26	308	308	R	1	--
27	309	311	R	3	--
28	312	326	R	15	7.20E+02
29	327	332	R	6	2.00E+01
30	333	344	K or R	12	3.46E+03
31	345	350	R	6	9.00E+00
32	351	359	K or Q	9	2.88E+02
33	360	367	R	8	6.40E+01

Serial No.	N-Term	C-Term	End With	Length	Variations
34	368	377	K	10	1.08E+02
35	378	391	K	14	4.80E+01
36	392	403	R	12	9.60E+02
37	404	413	R	10	2.70E+02
38	414	425	K	12	3.00E+00
39	426	433	K	8	9.00E+01
40	434	437	K or R	4	--
41	438	451	K	14	2.88E+04
42	452	470	K or R	19	9.22E+03
43	471	496	K	26	2.46E+04
44	497	515	R	19	3.69E+04
45	516	518	R	3	--
46	519	524	R	6	1.00E+00
47	525	536	R	12	1.73E+03
48	537	548	K	12	8.64E+02
49	549	572	R	24	8.19E+03
50	573	581	K	9	1.50E+01
51	582	584	R	3	--
52	585	590	K	6	3.20E+01
53	591	598	K	8	7.20E+01
54	599	602	R	4	--
55	603	620	K or R	18	1.23E+04
56	621	635	R	15	3.32E+05
57	636	641	R or S	6	6.40E+01
58	642	665	K	24	1.04E+04
59	666	674	K	9	9.60E+01
60	675	701	K	27	1.49E+06
61	702	704	T or R	3	--
62	705	721	R	17	8.64E+02
63	722	742	K	21	4.32E+04
64	743	746	K or R or P	4	--
65	747	749	A or R	3	--
66	750	757	G or R	8	1.20E+01
67	758	773	K	16	1.24E+05
68	774	797	P or R	24	3.89E+07
69	798	806	R	9	7.20E+01
70	807	809	R	3	--
71	810	823	K	14	1.08E+03
72	824	833	R	10	8.10E+02
73	834	853	R	20	2.30E+03
74	854	865	R	12	1.54E+03

Serial No.	N-Term	C-Term	End With	Length	Variations
75	866	872	K	7	2.40E+01
76	873	875	R	3	--
77	876	901	K or A	26	3.32E+05
78	902	905	R	4	--
79	906	923	R	18	7.37E+04
80	924	932	K or R	9	1.20E+01
81	933	935	K or R	3	--
82	936	944	R	9	8.00E+00
83	945	947	K	3	--
84	948	950	R	3	--
85	951	980	R	30	6.22E+04
86	981	991	K	11	7.20E+02
87	992	1007	R	16	9.60E+01
88	1008	1009	R or N	2	--
89	1010	1047	K or Q	38	6.19E+11
90	1048	1053	P or R	6	7.20E+02
91	1054	1056	K or R	3	--

Table 4.3.2 - Fragmentation for a2(I)

Serial No.	N-Term	C-Term	End With	Length	Variations
1	1	6	R	6	3.60E+01
2	7	21	R or P	15	6.80E+04
3	22	54	R or T or H	33	1.15E+10
4	55	62	K	8	1.50E+01
5	63	71	K or R or N	9	1.44E+02
6	72	74	K or R	3	--
7	75	78	R	4	--
8	79	87	R or A	9	2.16E+02
9	88	99	K	12	7.20E+01
10	100	102	K or R	3	--
11	103	111	K	9	9.60E+01
12	112	120	K	9	9.00E+02
13	121	138	R	18	6.05E+04
14	139	144	R	6	9.00E+00
15	145	146	R	2	--
16	147	156	R	10	7.20E+02

Serial No.	N-Term	C-Term	End With	Length	Variations
17	157	186	K	30	1.38E+06
18	187	204	R or P	18	1.87E+06
19	205	231	K	27	5.31E+06
20	232	249	R	18	2.30E+03
21	250	264	R	15	2.59E+03
22	265	276	K	12	2.40E+02
23	277	282	K	6	4.80E+01
24	283	302	K or R	20	4.67E+04
25	303	303	R	1	--
26	304	321	R	18	5.18E+05
27	322	327	R	6	2.40E+01
28	328	335	R	8	3.60E+01
29	336	345	R	10	1.08E+03
30	346	354	K or R	9	1.92E+03
31	355	362	R	8	2.70E+01
32	363	372	R	10	2.88E+02
33	373	386	K	14	2.16E+02
34	387	398	R	12	1.92E+03
35	399	408	R	10	7.20E+02
36	409	420	K	12	6.40E+01
37	421	428	K or E or Q	8	4.32E+02
38	429	432	K or R or T	4	--
39	433	441	R	9	8.64E+03
40	442	465	K	24	6.22E+06
41	466	491	K	26	4.61E+04
42	492	495	R or Q	4	--
43	496	510	K or R	15	2.30E+04
44	511	513	K or R	3	--
45	514	531	R	18	7.00E+05
46	532	543	K	12	1.01E+08
47	544	567	W or R	24	8.29E+06
48	568	576	K	9	5.40E+02
49	577	579	K	3	--
50	580	585	K or R or S	6	9.60E+02
51	586	593	K or R	8	1.12E+04
52	594	597	H or R	4	--
53	598	615	K or R	18	1.08E+05
54	616	630	R	15	1.15E+03
55	631	636	R	6	2.00E+01
56	637	660	K or R	24	1.56E+05
57	661	663	K or R	3	--

Serial No.	N-Term	C-Term	End With	Length	Variations
58	664	666	K or G or S or A	3	--
59	667	669	K	3	--
60	670	699	R or V	30	2.59E+11
61	700	716	R	17	2.16E+03
62	717	737	K	21	2.49E+05
63	738	741	R	4	--
64	742	744	R	3	--
65	745	752	R or A or P	8	3.00E+02
66	753	768	K	16	5.04E+04
67	769	801	R	33	4.20E+10
68	802	804	R or Q	3	--
69	805	828	R or I	24	2.59E+09
70	829	848	R or S	20	2.76E+07
71	849	860	R	12	2.40E+01
72	861	867	K	7	6.40E+01
73	868	870	R	3	--
74	871	896	K or R	26	1.24E+09
75	897	900	R	4	--
76	901	918	K or R	18	4.98E+06
77	919	927	M or R	9	9.60E+02
78	928	930	K	3	--
79	931	936	K or R or A	6	5.40E+01
80	937	939	R	3	--
81	940	945	K or R	6	1.20E+02
82	946	975	R	30	7.17E+07
83	976	986	K	11	4.00E+01
84	987	989	R or H or S	3	--
85	990	1002	R	13	9.72E+04
86	1003	1040	R	38	7.64E+08

Table 4.3.1 and Table 4.3.2 list all tryptic units from the alignments for both the α1(I) and α2(I) chains. The 'Start' and 'End' columns recorded the position of starting and ending residue for the relevant peptide fragment, while the 'End With' column shows the C-terminal residue for each peptide which is usually arginine (K) or lysine (R), and the 'Length' refers to the number of amino acid on that peptide chain. The 'Variations' Column shows the computing result of the variation number for peptides containing over 6 residues, and the algorithm will be explains later in Section 4.5.

4.4 Statistical Analysis on Variation

Based on the statistical analysis methodology demonstrated in *Section 3.2.3* and the result of sequence fragmentation shown above, variations of amino acid residues were analyzed for every tryptic peptide fragment listed in Table 4.3.1 and Table 4.3.2. The analysis was applied by counting and recording reside frequency in the alignment for each tryptic unit. As the record, such frequencies were assigned to the relevant matrix elements '$a_{i,j}$', so that series of matrices in the form of 'A' as illustrated in Formula 3.2.1 could be generated for the corresponding tryptic units.

The statistical analysis was carried out using 'R programme' in this research. R programme is a programming tool popularly used in the bioinformatics, the relevant codes are attached in appendix 4. As the outcome of the traversal of all tryptic units by computer programming, the statistical data for their residues variation can be calculated and stored in a series of numeric arrays as shown in Table 3.2.1. Consequently each tryptic peptide fragment has the statistical analysis data recorded in its specific matrix, as explained in the example from Section 2.2.4. Because the statistical analysis data which include 135 recording matrices could be too much to be illustrated in this thesis, an Excel file recorded main analyzing results in appendix7 and appendix 8.

4.5 Generation of Theoretical Sequence

On the basis of amino acid variation statistical analysis matrices, theoretical sequence can be generated according to the combination methodology demonstrated in Section 3.2.4. For each tryptic unit, all probable combinations of amino acid variations were generated to produce hypothetical peptide sequences

Taking G60 to K67 in α1(I) as an example, 48 possible combinations were produced for the alignment of this peptide as shown in Figure 4.5.1, where 48 is the product of the ‘Variations’ numbers shown in Table 4.5.1. Such hypothetical combinations were generated for each of the 70 α1 (I) and 65 α2 (I) tryptic units, and build up the database of UniColl as illustrated in Figure 4.5.2.

The amount of theoretical sequences generated for each tryptic unit has been computed using the above statistical method, and the results are shown in the ‘Variations’ columns in Table 4.3.1 and Table 4.3.2. All the calculation process and results are available in Appendix3, Appendix 7 and Appendix 8.

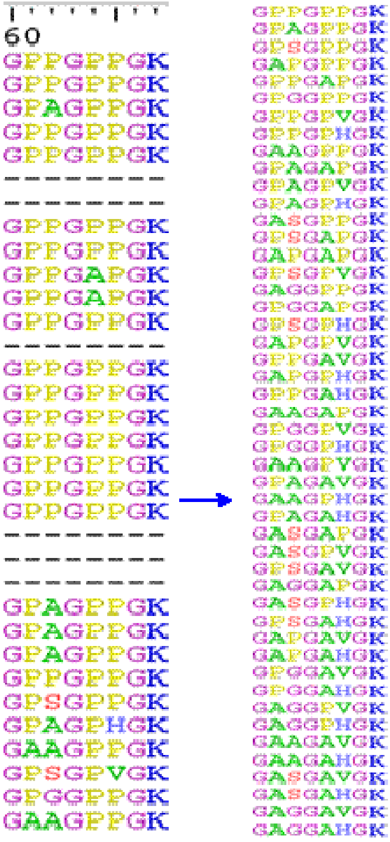


Figure 4.5.1 - The original alignment and theoretical peptides generated for COL1A1_0060-0067

The left column of the figure shows the original alignment of COL1A1_0060-0067 (the 60th to the 67th residues in α1(I) chain), and the right column shows the 48 theoretical peptides generated from this alignment using the above algorithm.

Table 4.5.1 - Amino acids variations exist in COL1A1_0060-0067

Position	Number of variations	Variations
60	1	G
61	2	P,A
62	4	P,A,S,G
63	1	G
64	2	P,A
65	3	P,H,V
66	1	G
67	1	K

The 'Position' column lists index numbers of the 60th to the 67th residues in the α1(I) chain, which corresponding to the nine amino acids in this tryptic unit. The middle column indicates the number of observed variations on each residue, while the variations are displayed in the last column.

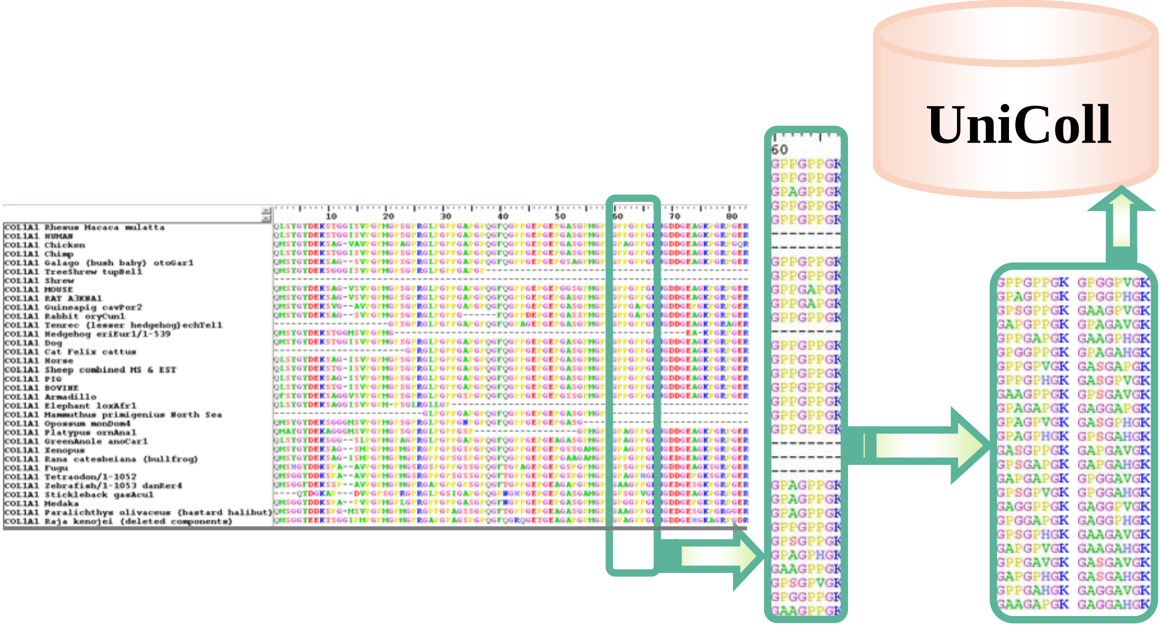


Figure 4.5.2 –Procedures of building UniColl

The above flow chart shows the generating process of UniColl. Hypothetical combinations are generated for all tryptic units in the alignment using the method demonstrated in Figure 4.5.1, then all combinations make up sequences in the database.

As explained in Section 3.2.5, the probability value of a theoretical tryptic peptide sequence could be calculated as the product of the probability of each amino acid in it (Formula 3.2.7), while the probability of an amino acid here could be computed as the frequency of its existence at that position through all species in the alignment.

Still take G60 to K67 in α1(I) as an example, the probability for hypothetical sequence 'GPPGPPGK' can be computed as: 1×(25/27) ×(17/27)×1×(25/27)×(25/27)×1×1=0.5.

The probability values for all theoretical sequence were calculated and recorded in Appendix 9 and Appendix 10 as in the 'score' columns.

4.6 Database Size Control

As shown in the 'Variations' columns in Table 4.3.1 and Table 4.3.2, the amount of theoretical sequences could be enormous for some of the peptide units that containing abundant variations. For example the peptide COL1A2_0621-0635 produced 2.59E+11 theoretical sequences.

Considering that too large a data capacity would reduce the efficiency and accuracy of the data search, especially in this case, a large part of the theoretical combinations could be very unlikely to occur (as evidenced by their low P values). Therefore a limitation was set up to cut down the database size. In this work, due to constraints placed by the MASCOT server, the size of UniColl was controlled to a manageable scale, with the threshold set as 1,000,000 theoretical seqeucnes for each tryptic unit. This means that the units generating more than one million of peptides were minimised by ignoring uncommon variations; the threshold can be adjusted in further research according to specific requirments.

Based on the database size control method explained in Section 3.2.7, uncommon variations were ignored in the tryptic units that produce too many combinations. Besides, if the number of theoretical peptides is still larger than one million after ignoring all the uncommon variations, the fish sequences could be excluded from the alignments since they were generally highly variable, and the UniColl will be mainly applied on mammal samples at this stage.

In the process of building UniColl, 30 tryptic units have been reduced by ignoring uncommon variations, and 8 of which have been reduced again by ignoring fish sequences. Since α2(I) chains are more variable than α1(I), the number of tryptic units which need to be cut due to excess of the threshold is bigger in α2(I) than in α1(I).

Table 4.6.1 - tryptic peptide units which have been reduced by cutting off uncommon variations and fish sequences

COL1A1 -uncommon	COL1A2 -uncommon	COL1A1 -uncommon; -fish	COL1A2 -uncommon; -fish
COL1A1_0027-0059	COL1A2_0007-0021	COL1A1_1010-1047	COL1A2_0022-0054
COL1A1_0210-0356	COL1A2_0121-0138		COL1A2_0544-0567
COL1A1_0497-0515	COL1A2_0157-0186		COL1A2_0670-0699
COL1A1_0621-0635	COL1A2_0283-0302		COL1A2_0769-0801
COL1A1_0675-0701	COL1A2_0442-0465		COL1A2_0871-0896
COL1A1_0722-0742	COL1A2_0466-0491		COL1A2_0946-0975
COL1A1_0774-0797	COL1A2_0496-0510		COL1A2_1003-1040
COL1A1_0951-0980	COL1A2_0514-0531		
	COL1A2_0598-0615		
	COL1A2_0637-0660		
	COL1A2_0717-0737		
	COL1A2_0805-0828		
	COL1A2_0829-0848		
	COL1A2_0901-0918		

'-uncommon' means infrequent variations in the alignment for the corresponding unit have been cut off; '-fish' mean fish sequences in the unit have been cut off.

4.7 Data Storage

According to the data storage rules demonstrated in Section 3.2.6, the theoretical peptide sequences were aligned into series of synthetic sequences, named by the index numbers of the beginning and ending residues that indicate their location in the collagen chain. As the result, 70 synthetic theoretical sequences for the α1(I) chain and 65 synthetic theoretical sequences for the α2(I) chain have been generated and listed in Table 4.7.1 and Table 4.7.2. Such sequences were fragmented into parts with length no more than 50000 amino acids to fit the standard for Mascot input. For example, if there are 80000 amino acids in the synthetic sequence of 'COL1A1_0010-0026', it will be cut into two secondary pieces in the database storage ('COL1A1_0010-0026-1' and 'COL1A1_0010-0026-2').

Table 4.7.1 - List of the 70 theoretic sequences generated from COL1A1

COL1A1_0001-0009	COL1A1_0255-0269	COL1A1_0471-0496	COL1A1_0758-0773
COL1A1_0010-0026	COL1A1_0270-0281	COL1A1_0497-0515	COL1A1_0774-0797
COL1A1_0027-0059	COL1A1_0282-0287	COL1A1_0519-0524	COL1A1_0798-0806
COL1A1_0060-0067	COL1A1_0288-0307	COL1A1_0525-0536	COL1A1_0810-0823
COL1A1_0068-0076	COL1A1_0312-0326	COL1A1_0537-0548	COL1A1_0824-0833
COL1A1_0077-0079	COL1A1_0327-0332	COL1A1_0549-0572	COL1A1_0834-0853
COL1A1_0084-0092	COL1A1_0333-0344	COL1A1_0573-0581	COL1A1_0854-0865
COL1A1_0093-0104	COL1A1_0345-0350	COL1A1_0585-0590	COL1A1_0866-0872
COL1A1_0108-0116	COL1A1_0351-0359	COL1A1_0591-0598	COL1A1_0876-0901
COL1A1_0117-0125	COL1A1_0360-0367	COL1A1_0603-0620	COL1A1_0906-0923
COL1A1_0126-0143	COL1A1_0368-0377	COL1A1_0621-0635	COL1A1_0924-0932
COL1A1_0144-0149	COL1A1_0378-0391	COL1A1_0636-0641	COL1A1_0936-0944
COL1A1_0152-0161	COL1A1_0392-0403	COL1A1_0642-0665	COL1A1_0951-0980
COL1A1_0162-0191	COL1A1_0404-0413	COL1A1_0666-0674	COL1A1_0981-0991
COL1A1_0192-0200	COL1A1_0414-0425	COL1A1_0675-0701	COL1A1_0992-1007
COL1A1_0201-0209	COL1A1_0426-0433	COL1A1_0705-0721	COL1A1_1010-1047
COL1A1_0210-0236	COL1A1_0438-0451	COL1A1_0722-0742	
COL1A1_0237-0254	COL1A1_0452-0470	COL1A1_0750-0757	

Table 4.7.2 - List of the 65 theoretic sequences generated from COL 1A2

COL1A2_0001-0006	COL1A2_0265-0276	COL1A2_0466-0491	COL1A2_0769-0801
COL1A2_0007-0021	COL1A2_0277-0282	COL1A2_0496-0510	COL1A2_0805-0828
COL1A2_0022-0054	COL1A2_0283-0302	COL1A2_0514-0531	COL1A2_0829-0848
COL1A2_0055-0062	COL1A2_0304-0321	COL1A2_0532-0543	COL1A2_0849-0860
COL1A2_0063-0071	COL1A2_0322-0327	COL1A2_0544-0567	COL1A2_0861-0867
COL1A2_0079-0087	COL1A2_0328-0335	COL1A2_0568-0576	COL1A2_0871-0896
COL1A2_0088-0099	COL1A2_0336-0345	COL1A2_0580-0585	COL1A2_0901-0918
COL1A2_0103-0111	COL1A2_0346-0354	COL1A2_0586-0593	COL1A2_0919-0927
COL1A2_0112-0120	COL1A2_0355-0362	COL1A2_0598-0615	COL1A2_0931-0936
COL1A2_0121-0138	COL1A2_0363-0372	COL1A2_0616-0630	COL1A2_0940-0945
COL1A2_0139-0144	COL1A2_0373-0386	COL1A2_0631-0636	COL1A2_0946-0975
COL1A2_0147-0156	COL1A2_0387-0398	COL1A2_0637-0660	COL1A2_0976-0986
COL1A2_0157-0186	COL1A2_0399-0408	COL1A2_0670-0699	COL1A2_0990-1002
COL1A2_0187-0204	COL1A2_0409-0420	COL1A2_0700-0716	COL1A2_1003-1040
COL1A2_0205-0231	COL1A2_0421-0428	COL1A2_0717-0737	
COL1A2_0232-0249	COL1A2_0433-0441	COL1A2_0745-0752	
COL1A2_0250-0264	COL1A2_0442-0465	COL1A2_0753-0768	

These sequences were integrated into two FASTA files respectively for α1(I) and α2(I), which composing the UniColl database. The copies of txt files converted from these two FASTA files can be found in the Appendix 5 and Appendix 6; while some more detailed information of data in UniColl were illustrated in Appendix 9 and Appendix 10, which listed the top 1000 theoretical combinations for each tryptic unit, under the order of probability value from high to low.

4.8 Conclusion

As the statistic result, data inputted to Unicoll included approximately 3.0×10^6 synthetic sequences for α1(I) and 1.4×10^6 ones for α2(I), which covered almost all probable combinations generated from the known type I collagen sequences, except a small number of very unlikely ones listed in Table 4.6.1. Joining of these tryptic units together can produce full type I collagen sequences with the equivalent of 4.9×10^{211} possible combinations for α1(I) and 1.1×10^{211} ones for α2(I). The size of UniColl is significantly larger than any known protein database.

As the sequences in UniColl are aligned in groups of tryptic units other than in species, UniColl is a peptide-specific database, a basic difference from most other protein databases.

Chapter 5

Database Evaluation

Evaluation on the database was carried out in order to test and estimate the feasibility and effectiveness of UniColl. As the result, the practical utility and high quality results of this database have been proved in the test.

5.1 Methodology

5.1.1 Researching Procedure

The evaluating procedure is demonstrated as in Figure 5.1.1, with the steps of noise filtering and calibration applied on the experimental sample, database searching was applied and the searching results were analysed in comparison with other databases from the two criteria of coverage and accuracy.

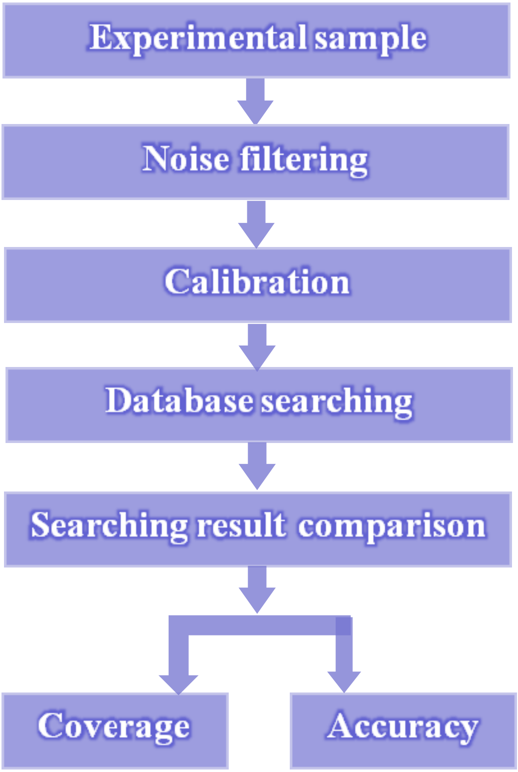


Figure 5.1.1 - The procedure of evaluating UniColl

5.1.2 Database Searching

Mass spectra of both fresh and fossil bone collagen samples were searched against the three databases for comparison. Mascot was applied as the searching engine in this test. It provides reports of searching results, with information of matched proteins and peptides contained.

5.1.3 Scoring System

Mascot is based on the Mowse scoring algorithm (Pappin *et al.* 1993) to assign MS/MS spectra to peptides. The mass-to-charge ratio of experimental MS/MS fragment ions are matched to calculated masses for each entry from the sequence database on a probabilistic basis. Matches are judged by the reported ion score calculated as '$-10 * \text{Log}(P)$', where 'P' is the absolute probability that the observed match is random. Hereafter peptide matches with high scores have low probability of being random matches.

5.1.4 Mapping for Sequence

From the peptide matches showed on mascot reports, samples were sequenced and mapped as proteins. Sequences of α1(I) and α2(I) chains for different species from different databases were generated and mapped into four 'fasta' files respectively for human, cow, dodo and giant tortoise (see Appendix 7 to 10). These mappings were then be estimated by their coverage and accuracy.

5.2 Materials

Fresh bone collagen samples:

Human and cow mass spectrometry data (H. Koon, pers.comm, August 2007.) were applied as the materials for test of UniColl. Seven tandem MS peak lists for bone collagen of human and another seven for cow were applied.

Ancient bone collagen samples:

MS/MS data of fossil bone collagen samples from Mauritius for extinct dodo (Hume 2006) and giant tortoise were the other two examples in our test.

5.3 Databases for comparison

UniProt:

'UniProt', the universal protein database, is the integration of the three popular protein databases of Swiss-Prot, TrEMBL and PIR, is one of the main public species-specific resources of protein data. However type I collagen sequences are very limited in UniProt now, only seven species were fully covered for either α1(I) or α2(I) chains.

Collagens:

'Collagens' is a species-specific collagen database composed of type I collagen sequences from the known 40 species. This collagen-specific database, covering most information of type I collagen in hand, was developed in house combined with sequences from UCSC and genomic data.

UniColl:

'UniColl' is a novel theoretical type I collagen sequence database containing almost all known probable theoretical sequences for α1(I) and α2(I) chains. Approximately 3.0×10^6 for α1(I) and 1.4×10^6 for α2(I) hypothetical peptide fragments were contained in UniColl. Different to UniProt and Collagen, UniColl is a peptide-specific collagen database.

5.4 Criteria for Database Evaluation

5.4.1 Coverage

The coverage of peptides matched in protein sequences is an important criterion for evaluating protein database searching. In this test, the coverage of sequences was described by the number of matched amino acids, peptides, and the length of those peptides.

In order to determine valid hits of peptides matches, matches with low ion scores were excluded from the hits assigned to proteins. The decoy database abstracted matches of false-positive which could not be included in valid hits. With the assistance of decoy, all matches with scores lower than the maximum of false-positive score were re-examined. This examination was based on quality of experimental spectra, number of ion mass matches, and type of matched ions.

5.4.2 Accuracy

The accuracy of covered sequences is another important criterion for judging database searching results. Here the accuracy was measured by the number of peptides that were misidentified, and the number of amino acids which caused the misidentification.

The misidentification is defined as the differences in matching result between the true protein sequence and the sequence which peptide is identified with the highest score. Therefore, the results for type I collagen of human and cow can be evaluated by accuracy, because the true sequences are known. The accuracy estimation on fresh samples will be applied to test the ability of UniColl to identify true sequences. This is the main purpose of the fresh sample experiment.

For ancient samples such as dodo and giant tortoise, it is not practical to evaluate accuracy, since their sequences are not complete. The accuracy of assessment of an unknown sample could refer to the quality of MS/MS match. The purpose of the experiment on ancient samples was to test the ability of UniColl to identify unknown samples and to identify peptides not found in conventional databases.

5.5 Results for Fresh Samples

The fresh samples were tested first, because their sequences presented in all of the three databases, ensured that all high quality spectra could be matched to corresponding type I collagen sequences in databases. Therefore the probability of misidentification caused by absence of sequences in databases can be ignored. The results were then compared to ancient sample whose sequences were not covered in any of the three databases.

5.5.1 Coverage for Fresh Samples

As the result of fresh samples database searching, 1416 peptides (783 for COL1A1 and 633 for COL1A2) and the equivalent of 26405 amino acid residues were covered in the searching outcomes; the distribution of hits in the three databases is shown in Table 5.5.1 and Table 5.5.2.

From the view of the two investigated species, 36 α1(I) and 30 α2(I) tryptic peptides for human, 43 α1(I) and 36 α2(I) ones for cow were covered. In these peptides, 29 α1(I) and 29 α2(I) for human overlapped in three databases, while 34 α1(I) and 34 α2(I) for cow overlapped (Table 5.5.3). That means the searching results for three databases had around 87% in common, while UniColl obtained 6 unique peptide hits compared to the other two databases that covered slightly more than UniColl. The matched peptides for fresh samples are listed in Appendix 7 and Appendix 8.

Table 5.5.1 - Nnumber of peptides covered for fresh samples.

Peptides	Total (1416)	COL1A1 (783)	COL1A2 (633)	Human COL1A1	Human COL1A2	Cow COL1A1	Cow COL1A2
UniColl	468	254	214	100	81	154	133
UniProt	476	266	210	102	77	164	133
Collagens	472	263	209	102	76	161	133

Table 5.5.2 - Number of amino acids covered for fresh samples.

Amino acids	Total (26405)	COL1A1 (14380)	COL1A2 (12025)	Human COL1A1	Human COL1A2	Cow COL1A1	Cow COL1A2
UniColl	8830	4746	4084	1941	1590	2805	2494
UniProt	8825	4848	3977	1944	1511	2904	2466
Collagens	8750	4786	3964	1944	1497	2842	2467

Table 5.5.3 - Number of tryptic units covered for fresh samples.

Tryptic units	Total (145)	COL1A1 (79)	COL1A2 (66)	Human COL1A1 (36)	Human COL1A2 (30)	Cow COL1A1 (43)	Cow COL1A2 (36)
UniColl	133	68	65	31	30	37	35
UniProt	139	75	64	34	29	41	35
Collagens	139	75	64	34	29	41	35

Table 5.5.1-5.5.3 display the number of hits for peptides or residues of fresh sample database searching for the three databases listed in the first columns. Numbers in tables indicate the number of hits for corresponding classes in the three database, while numbers in brackets indicate the total number of hits for the corresponding class, for example 'COL1A1 (14380)' in Table 5.5.2 means there are 14380 amino acids covered for the α1(I) chain.

For each database, specific hits existed for various reasons. The unique peptide hits (shown in Table 5.5.4) for 'UniProt' and 'Collagens' found in this investigation were mainly from missed cleavages of fragments, which cannot be identified in UniColl. One exception is the telopeptide COL1A1_0010-0026 for cow, which contains an absent ‘G’, which can only be identified by UniProt. The results reveal the difficulty of identifying missed cleavage in UniColl, and the problem of identifying amino acids absence in both UniColl and Collagen.

Table 5.5.4 - Peptides covered in UniProt and Collagen but not UniColl.

Peptide index	Peptide sequence
COL1A1_0360-0377	GSPGEAGRPGEAGLPGAK
COL1A1_0392-0413	TGPPGPAGQDGRPGPPGPPGAR
COL1A1_0471-0515	GEQGPAGSPGFQGLPGPAGPPGEAGKPGEQGVPGDLGAPGPSGAR
COL1A1_0150-0161	GRPGAPGPAGAR
COL1A1_0750-0773	GETGPAGRPGEVGPPGPPGPAGEK
COL1A2_0466-0495	GEQGPAGPPGFQGLPGPAGTAGEAGKPGER
COL1A1_0010-0026	STG~ISVPGPMGPSGPR

In the above table, the first six peptides are miss-cleaved, while the last one contains the absence of one amino acid.

Peptides hits specific in UniColl (shown in Table 5.5.5) covered four tryptic units, three of which come from ‘Cow 20070124’, suggesting the high quality of this sample set; while the other unique hit presented in most of the fourteen samples.

Table 5.5.5 - Unique peptides matched in UniColl.

Peptide index	Peptide sequence	Sample set
COL1A1_0001-0009	QLSYGYDEK	Cow 20070124
COL1A1_0471-0496	GEQGPAGPPGFQGLPGPAGAAGETGK	Cow 20070124
COL1A1_0824-0833	QGPSGASGER	Cow 20070124
COL1A2_0387-0398	EGPVGLPGIDGR	Most human & cow

In the above table, the first three peptides are unique hit in UniColl from cow, while the last one peptide exists in most of the fourteen samples.

Short peptides with length of four or less were not covered in any of the three databases, except two of which included in the miss-cleaved fragments. This shows that short fragments are difficult to be identified even in databases that contain them, also proves that it is reasonable to exclude short tryptic fragments in building of UniColl. It shows that the type of MALDI mass-spectrometry we did is very poor at identifying short peptides due to the mass of matrix peaks.

In conclusion, compared to the mascot matching results of other databases, UniColl has the ability to get MS/MS spectra matched to sequences from known species, except it is unable to identify missed cleavage and absence in peptides. UniColl also has the ability to identify peptides that are not recognized in other databases.

5.5.2 Accuracy for Fresh Samples

In searching on species-specific databases UniProt and Collagen, all fourteen type I collagen samples from either human or cow were identified as the corresponding species. Since full sequences of COL1A1 and COL1A2 for these two species are included in databases, which assigned peptide matches to protein hits, the matches selected for given species are 100% identical to the real sequence.

However in the peptide-specific database UniColl, peptide matches are not assigned to proteins but tryptic peptide fragments marked with their position. The most likely match for each peptide is selected by its ion score compared with each other alternative match for that unit. Accordingly, such peptide units located in the result protein sequence are probably not from a unique species.

As the result in UniColl, several peptides are misidentified as collagen sequences not from the true species. From the test on seven type I collagen samples of human, 16 peptides out of 31 matched ones for α1(I) and 12 out of 30 for α2(I) include misidentifications. While for cow samples, 18 peptides out of 37 hits for α1(I) and 20 out of 35 hits for α2(I) contain problems in matching to proper species. As shown in Table 5.5.6, the mean accuracy rate of UniColl is 50.7%, which is only half of the accuracy rate of the other two databases which is almost 100%.

Table 5.5.6 - Accuracy of UniColl database searching

Sample set	Human-α1(I)	Human-α2(I)	Cow-α1(I)	Cow-α2(I)
Number of matches	31	30	37	35
Number of accurate matches	15	18	19	15
Accuracy	48.4%	60%	51.4%	42.9%

The above table shows the numbers of matched peptide in UniColl and the accuracy rate of these matching. The accuracy value is proportion of the number of accurate matches to the number of matches.

When examined the misidentifications, most of them came from poor spectra. Noise peaks and absence of ions would generate difficulties in the matching, and they can only match accurately in the 'Collagens' database with no other equally plausible peptides present. Fragments misidentified of this reason were mainly corresponding to weak spectra. Some other misidentified spectra that included enough information but were assigned to sequences other than the 'right answer'. Considering that they had higher ion scores in matched peptides than the 'right answer', this revealed the probability that these spectra came from contamination.

Accordingly, compared to the mascot matching results of other databases, UniColl is more likely to generate misidentifications. The reason for this is UniColl contains far more sequences selections for MS/MS peptide matching than any other conventional collagen databases, induces a higher false positive rate of matching. However, these misidentifications could reveal the existence of poor spectra, with insufficient ion fragments to indentify the true sequences.

5.6 Results for Ancient Samples

5.6.1 Coverage for Ancient Samples

UniColl is expected to produce higher coverage than the other two databases in ancient sample identification, as ancient sequences are not covered in databases of UniProt and Collagens, but UniColl includes all hypothetical type I collagen sequences that are more possible to cover ancient sequences.

In fact UniColl does cover more peptides than UniProt and Collagens in the test for dodo and giant tortoise (proved by data shown in Table 5.6.1, Table 5.6.2 and Figure 5.6.1, Figure 5.6.2). In our dodo sample, two novel α1(I) and two α2(I) peptides were covered, while four α1(I) and five α2(I) fragments in giant tortoise were found. These peptides with differences to known sequences match to their MS/MS spectra well.

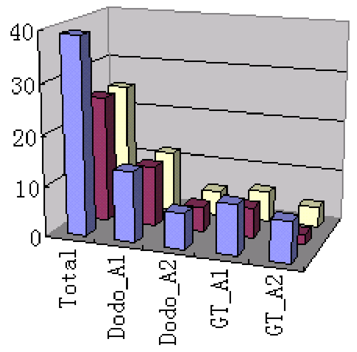
Table 5.6.1 - Peptides covered for ancient samples.

Peptides	Total (91)	COL1A1 (60)	COL1A2 (31)	Dodo COL1A1	Dodo COL1A2	GT COL1A1	GT COL1A2
UniColl	39	24	15	14	7	10	8
UniProt	25	18	7	12	5	6	2
Collagen	25	18	7	12	5	6	4

Table 5.6.2 - Amino acids covered for ancient samples.

Amino acids	Total (1868)	COL1A1 (1239)	COL1A2 (629)	Dodo COL1A1	Dodo COL1A2	GT COL1A1	GT COL1A2
UniColl	814	503	311	285	146	218	165
UniProt	506	368	138	243	105	125	33
Collagen	548	368	180	243	105	125	75

Table 5.6.1 and Table 5.6.2 show the number of hits for peptides or residues of ancient sample database searching for the three databases listed in the first columns. Numbers in tables indicate the number of hits for corresponding classes in the three database, while numbers in brackets indicate the total number of hits for the corresponding class, for example 'COL1A1 (60)' in Table 5.6.1 means 60 peptides are matched for the α1(I) chain.



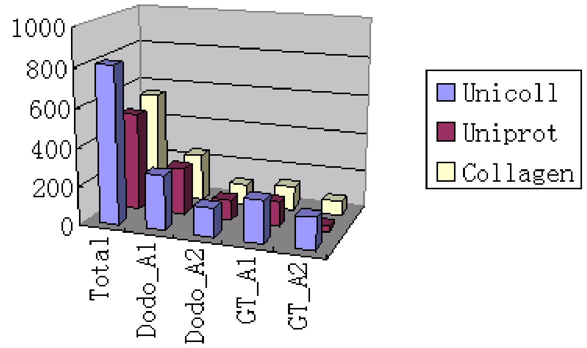


Figure 5.6.1- Number of peptides covered *Figure 5.6.2- Number of amino acids covered*

The above figures display the distribution of ancient samples database searching results, while Figure 5.6.1 concerns the number of peptide hits, and Figure 5.6.2 concerns the number of amino acid hits. In each figure, the 'Total' column indicates the total number of hits in the three databases; the other four columns show the number of hits separately for α 1(I) or α 2(I) chains of dodo or giant tortoise. The x-axis is the content of columns, and y-axis is the number of matches.

In peptides identified from UniColl, thirteen novel sequences were obtained (Table 5.6.3), in which 4 sequences came from dodo, and 9 came from giant tortoise. These peptides were not recognized in UniProt and Collagens, because the sequences were not included. The advantage of huge sequence source in UniColl was proved.

Table 5.6.3 - Novel coverage for dodo and giant tortoise in UniColl

Peptide index	Peptide mass	Ion score	Peptide sequence
DODO COL1A1_0549-0572	2197.1809	119	GDAGAPGAPGGEGPPGLEGMPGER
DODO COL1A1_0906-0923	1596.8972	53	GEPGPAGPPGPIGPAGPR
DODO COL1A2_0717-0737	1828.0215	69	VGPPGPAGISGPSGLPGPPGK
DODO COL1A2_0829-0848	1816.9436	104	GPPGPIGMPGLAGPPGEAGR
GT COL1A1_0162-0191	2547.2903	95	GNDGAVGAAGPPGPTGPAGPPGFPGAVGAK
GT COL1A1_0471-0496	2339.1453	71	GEQGIAGAPGFQGLPGPAGAPGEAGK
GT COL1A1_0705-0721	1529.7771	135	GNAGPPGPTGFPGAAGR
GT COL1A1_0834-0853	1724.8849	126	GPPGPAGPPGLAGPPGEAGR
GT COL1A2_0205-0231	2319.2295	72	GEIGLPGASGPVGPAGNPGANGLAGAK
GT COL1A2_0466-0491	2339.1453	71	GEQGPAGAPGFQGLPGPAGAPGEAGK
GT COL1A2_0700-0716	1529.7771	135	GDAGPPGLTGFPGAAGR
GT COL1A2_0805-0828	2136.9858	113	GLPGISGGNGEPGPAGISGPSGPR
GT COL1A2_0829-0848	1740.8785	90	GPPGAIGPPGLAGPPGEAGR

As shown in Table 5.6.3, most matches have very high ion scores, many of which are over one hundred. This suggests the reliability of these matches, proving the novel hits in UniColl. An example of the MS/MS matching in three databases for the peptide mass of 2547.2903 in giant tortoise sample is shown in Figure 5.6.1 to 5.6.3 to prove.

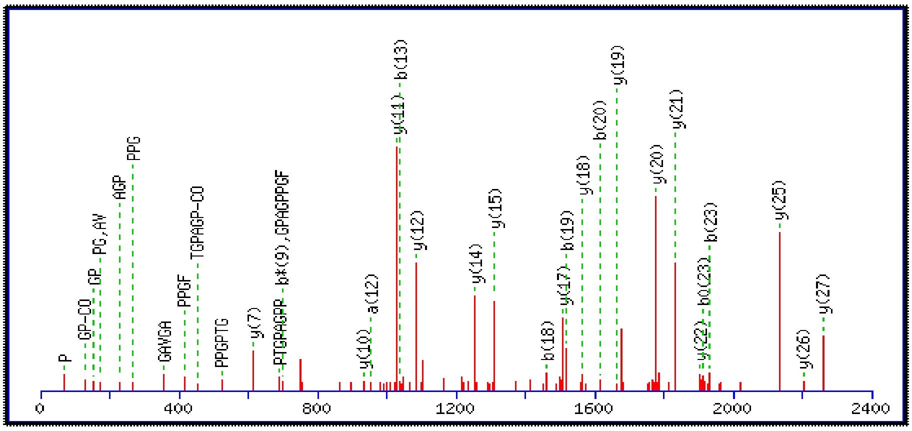


Figure 5.6.3 - MS/MS spectrum for peptide 2547.2903 matching to sequence 'GNDGAVGAAGPPGPTGPAGPPGFPGAVGAK' in UniColl with ion score of 95.

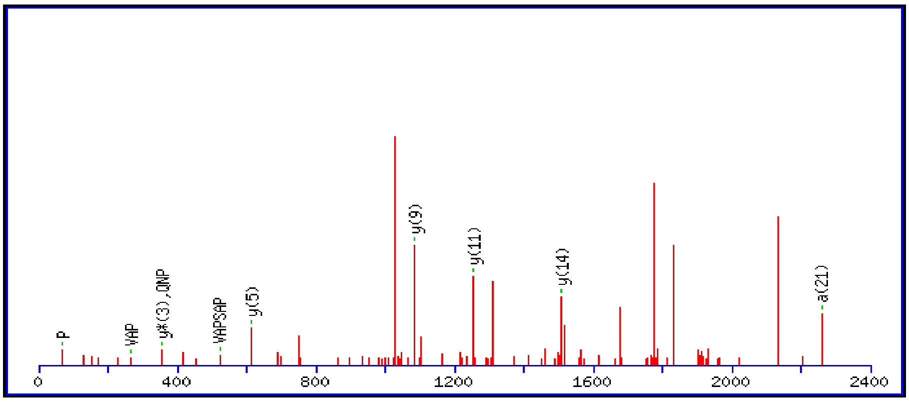


Figure 5.6.4 - MS/MS spectrum for peptide 2547.2903 matching to sequence 'YIIHVPTVVAPSAPIFNPQNPLK' in UniProt with ion score of 8.

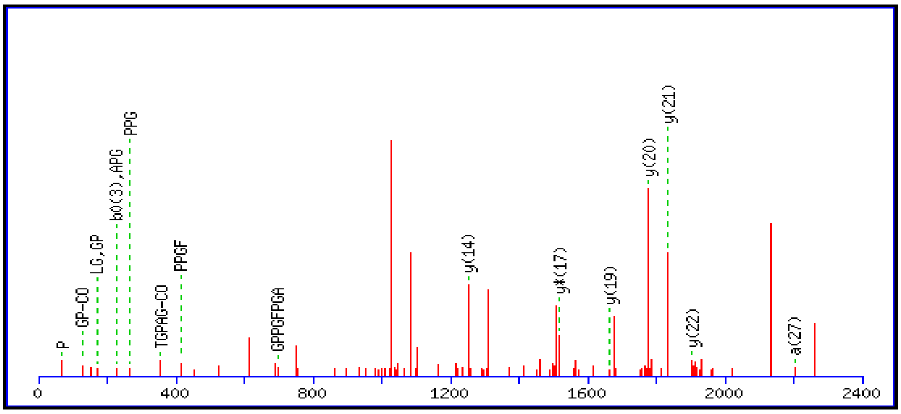


Figure 5.6.5 - MS/MS spectrum for peptide 2547.2903 matching to sequence 'GADGSTGPAGPAGPLGAAGPPGFPGAPGPK' in Collagen with ion score of 1.

In Figure 5.6.1, the illustration for MS/MS database matching for peptide 2547.2903 in UniColl, almost all y ions m/z values are assigned to peaks in spectrum, giving a high ion score of 95 matching to 'GNDGAVGAAGPPGPTGPAGPPGFPGAVGAK'. However as the results of same peptide in the other two databases, most main peaks are not matched to any fragment ions. In UniProt, the peptide is identified as a non-collagen sequence with the ion score of 8, while in Collagens it is matched with ions score of only 1. The reason why a good match is not available in these two databases is that none of the sequences in them can match to the target spectrum. While the matching result of this spectrum from UniColl is reliable due to the high ion score.

5.6.2 Accuracy for Ancient Samples

As discussed before, the accuracy evaluation for ancient sample database searching cannot be precisely estimated, since the 'right' sequences have not been defined. However the sequences matched for dodo were close to chicken type I collagen, suggesting the good accuracy in matching. The accuracy of Giant tortoise sequences was more difficult to examine, since no congeneric species was included in database.

According to the difficulty in evaluating the accuracy of matches directly, an investigation was carried out for peptide that got different matches in UniColl and the other two databases; and to judge which is the better match from their spectra. Groups of such peptides are shown below for dodo (Table 5.6.4) and giant tortoise (Table 5.6.5) separately.

Table 5.6.4 - dodo type I collagen peptide matches comparison within different databases.

Peptide index	Peptide sequence matched

COL1A1_0312-0326	GEPGPAGPPGSPGER(UniColl)
	GEPGPAGLPGPAGER(UniProt& Collagens)

COL1A1_0642-0665	GEPGLPGPAGFAGPPGADGQPGAK(UniColl)
	GEPGPPGPAGFAGPPGADGQPGAK(UniProt& Collagens)

COL1A1_0722-0742	VGPPGPAGNIGLPGPPGPAGK(UniColl)
	VGPPGPSGNIGLPGPPGPAGK(UniProt& Collagens)

COL1A1_0774-0797	GSPGADGPPGAPGTPGPQGIAGQR(UniColl)
	GSPGADGPIGAPGTPGPQGIAGQR(UniProt& Collagens)

COL1A1_0951-0980	GFSGLQGPPGPPGSPGEQGPAGASGPAGPR(UniColl)
	GFSGLQGPPGPPGAPGEQGPSGASGPAGPR(UniProt& Collagens)

COL1A2_0022-0054	GPPGPPGPPGPQGFQGPPGEPGEPGQTGPQGPR(UniColl)
	GPPGASGPPGPPGFQGVPGEPGEPGQTGPQGPR(UniProt& Collagens)

COL1A2_0637-0660	GEPGPVGPSGFAGPPGAAGQSGPK(UniColl)
	GEPGPVGPSGFAGPPGAAGQPGAK(UniProt& Collagens)

Table 5.6.5 - giant tortoise type I collagen peptide matches comparison within different databases.

Peptide index	Peptide sequence in matched

COL1A1_0549-0572	GDAGAAGNPGNQGPPGLQGMPGER(UniColl)
	GDAGAPGAPGNEGPPGLEGMPGER(UniProt& Collagens)

COL1A1_0642-0665	GEPGAVGHAGFAGPPGADGQPGAK(UniColl)
	GEPGPPGPAGFAGPPGADGQPGAK(UniProt& Collagens)

COL1A1_0722-0742	VGPPGPAGNIGLPGPPGPSGK(UniColl)
	VGPPGPSGNIGLPGPPGPAGK(UniProt& Collagens)

COL1A1_0951-0980	GFSGLQGLPGPAGPPGEQGPSGASGPAGPR(UniColl)
	GFSGLQGPPGPPGAPGEQGPSGASGPAGPR(UniProt& Collagens)

COL1A2_0496-0510	GIPGEFGLPGLAGPR(UniColl)
	GLPGEFGLPGPAGPR(UniProt& Collagens)

COL1A2_0901-0918	GEPGPTGPTGPVGPAGAR(UniColl)
	GEPGPAGSIGPVGAAGPR(Collagens)

In this investigation of peptides from Table 5.6.4 and Table 5.6.5, most matches in UniColl were found better than the ones in other databases either in ion scores or from the spectra view. An example is shown in Figure 5.6.4 and 5.6.5 as follows.

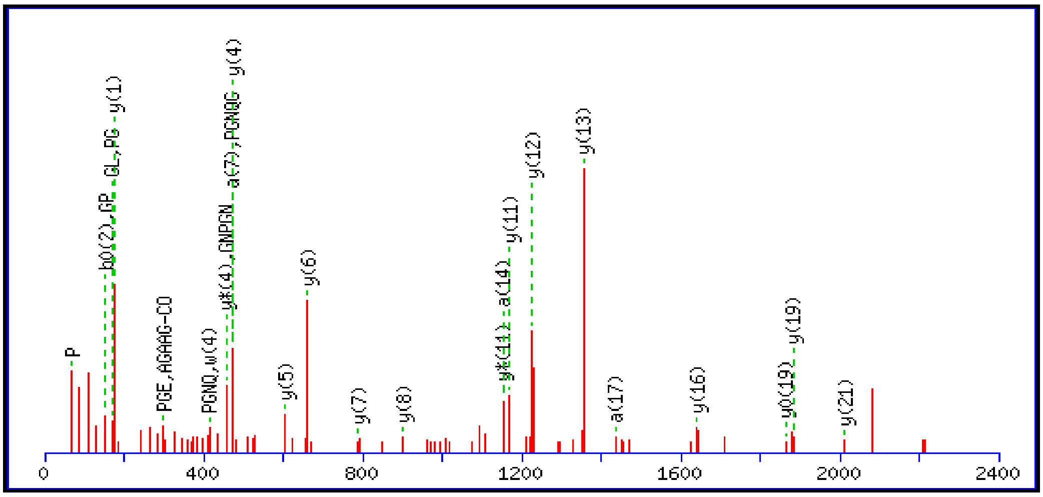


Figure 5.6.6 - MS/MS spectrum for peptide 2255.0979 matching to sequence ‘GDAGAAGNPGNQGPPGLQGMPGER’ in UniColl (ion score:60).

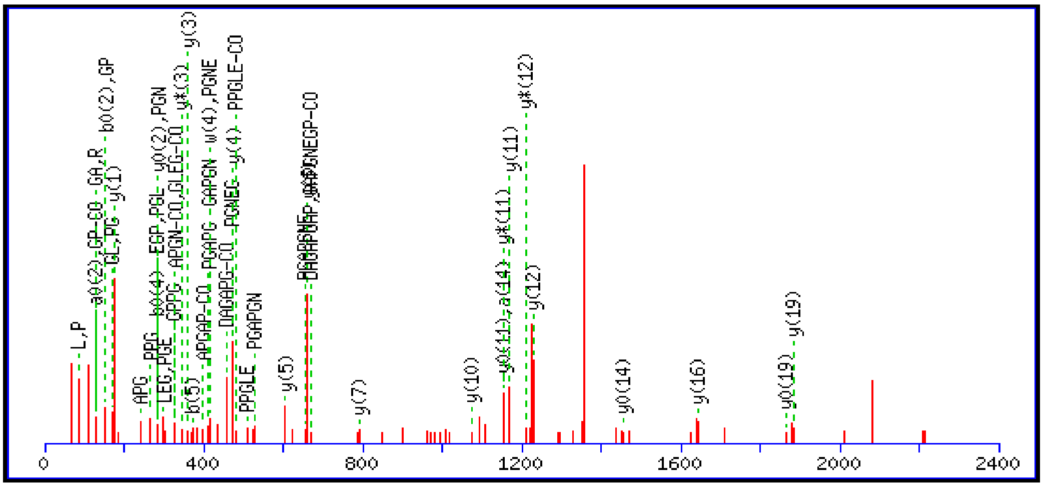


Figure 5.6.7 - MS/MS spectrum for peptide 2255.0979 matching to sequence ‘GDAGAPGAPGNEGPPGLEGMPGER’ in UniProt and Collagens (ion score: 38).

Figure 5.6.4 shows the MS/MS spectrum for peptide 2255.0979 from giant tortoise, matching to sequence 'GDAGAAGNPGNQGPPGLQGMPGER' in UniColl, while Figure 5.6.5 shows the same peptide but in different databases of UniProt and Collagen that matched to 'GDAGAPGAPGNEGPPGLEGMPGER'. There are four amino acids in different between these two sequences, however the match in UniColl has a higher ion score and more main peaks assigned to y ions, especially the main peak of y(13), than the matches in UniProt and Collagen, which suggest that the UniColl match is more accurate. More interestingly, there is a unique sequence QXXXXXQ (X stands for other amino acids) pattern in the UniColl sequence, which is identical to the relevant peptide sequence of the green anole, another reptile.

Similar patterns specific between dodo and giant tortoise were also found in the matches from UniColl, might reveal relationship of this two ancient species that were living together.

5.7 Publications

The UniColl has been applied as a powerful tool to discover novel sequences from ancient samples in BioArch laboratory, York University. Some of the achievements have been published, and following are two publications with contribution of UniColl.

5.7.1 Publication I

The following paper was published on *Science* in 2008:

Mike Buckley, Angela Walker, Simon Y. W. Ho, **Yue Yang**, Colin Smith, Peter Ashton, Jane Thomas Oates, Enrico Cappellini, Hannah Koon, Kirsty Penkman, Ben Elsworth, Dave Ashford, Caroline Solazzo, Phillip Andrews, John Strahler, Beth Shapiro, Peggy Ostrom, Hasand Gandhi, Webb Miller, Brian Raney, Maria Ines Zylber, M. Thomas P. Gilbert, Richard V. Prigodich, Michael Ryan, Kenneth F. Rijsdijk, Anwar Janoo and Matthew J. Collins, "Comment on 'Protein Sequences from Mastodon and Tyrannosaurus rex Revealed by Mass Spectrometry' ", *Science*, 319(2008) 33

In this paper, UniColl was used to test the *Tyrannosaurus rex* collagen peptide sequences claimed by Asara *et al.* in 2007. As the result of running their MS data in UniColl, collagen sequences other than the claimed ones have been recovered with higher matching rate to the fossil sample. The result suggested that the claimed T. rex a1(I) collagen sequences were not necessarily the best answer can be discovered from the fossils. Meanwhile, UniColl provides more possibilities to identify unknown samples such as ancient fossils, of which sequences were not included in any other protein databases.

The full text is supplemented as follows.

Science AAAS

Comment on "Protein Sequences from Mastodon and Tyrannosaurus rex Revealed by Mass Spectrometry"

Mike Buckley, *et al.*

Science **319**, 33c (2008);
DOI: 10.1126/science.1147046

Comment on "Protein Sequences from Mastodon and *Tyrannosaurus rex* Revealed by Mass Spectrometry"

Mike Buckley,[1] Angela Walker,[2] Simon Y. W. Ho,[3] Yue Yang,[1] Colin Smith,[4] Peter Ashton,[1] Jane Thomas Oates,[1] Enrico Cappellini,[1] Hannah Koon,[1] Kirsty Penkman,[1] Ben Elsworth,[1] Dave Ashford,[1] Caroline Solazzo,[1] Phillip Andrews,[2] John Strahler,[2] Beth Shapiro,[6] Peggy Ostrom,[5] Hasand Gandhi,[5] Webb Miller,[6] Brian Raney,[7] Maria Ines Zylber,[8] M. Thomas P. Gilbert,[9] Richard V. Prigodich,[10] Michael Ryan,[11] Kenneth F. Rijsdijk,[12] Anwar Janoo,[13] Matthew J. Collins[1]*

We used authentication tests developed for ancient DNA to evaluate claims by Asara *et al*. (Reports, 13 April 2007, p. 280) of collagen peptide sequences recovered from mastodon and *Tyrannosaurus rex* fossils. Although the mastodon samples pass these tests, absence of amino acid composition data, lack of evidence for peptide deamidation, and association of α1(I) collagen sequences with amphibians rather than birds suggest that *T. rex* does not.

Early reports of DNA preservation in multimillion-year-old bones (i.e., from dinosaurs) have been largely dismissed (*1*, *2*) (table S1), but reports of protein recovery are persistent [see (*3*) for review]. Most of these studies used secondary methods of detection, but Asara *et al*. (*2*) recently reported the direct identification of protein sequences, arguably the gold standard for molecular palaeontology, from fossil bones of an extinct mastodon and *Tyrannosaurus rex*. After initial optimism generated by reports of dinosaur DNA, there has been increasing awareness of the problems and pitfalls that bedevil analysis of ancient samples (*1*), leading to a series of recommendations for future analysis (*1*, *4*). As yet, there are no equivalent standards for fossil protein, so here we apply the recommended tests for DNA (*4*) to the authentication of the reported mastodon and *T. rex* protein sequences (*2*) (Table 1).

First, the likelihood of collagen survival needs to be considered. The extremely hierarchical structure of collagen results in unusual, catastrophic degradation (*5*) as a consequence of fibril collapse. The rate of collagen degradation in bone is slow because the mineral "locks" the components of the matrix together, preventing helical expansion, which is a prerequisite of fibril collapse (*6*). The packing that stabilizes collagen fibrils (*6*) also increases the temperature sensitivity of degradation (E_a 173 kJ mol^{-1}) (Fig. 1). Collagen decomposition would be much faster in the *T. rex* buried in the then-megathermal (>20°C) (*7*) environment of the Hell Creek formation [collagen half-life ($T_{1/2}$) = ~ 2 thousand years (ky] than it would have been in the mastodon lying within the Doeden Gravel Beds (present-day mean temperature, 7.5°C; collagen $T_{1/2}$ = 130 ky) (Fig. 1).

This risk of contamination also needs to be evaluated. Collagen is an ideal molecular target for this assessment because the protein has a highly characteristic motif that is also sufficiently variable to enable meaningful comparison between distant taxa if enough sequence is obtained (Fig. 2). Compared with ancient DNA amplification, contamination by collagen is inherently less likely. Furthermore, because the bones sampled in (*2*) were excavated by the

Table 1. Key questions to ask about ancient biomolecular investigations [adapted from (*4*)].

Test	Sample	Pass	Observation
Do the age, environmental history, and preservation of the sample suggest collagen survival?	Mastodon, 300 to 600 ky old	Yes	Collagen $T_{1/2}$ at 7.5°C = 130 ky
	T. rex, 65 million years old	No	Collagen $T_{1/2}$ at 20°C = 2 ky
Do the biomolecular and/or macromolecular preservation of the sample, the molecular target, the innate nature of the sample, and its handling history suggest that contamination is a risk?	Biomolecular preservation	?	Range of evidence presented (*8*) but no amino acid compositional data
	Macromolecular preservation	Yes	Macromolecular preservation is not the equivalent of biomolecular preservation (*9*)
	Molecular target	Yes	
	Handling history	Yes?	Large (2.5 g) samples increase risk of contamination?
Do the data suggest that the sequence is authentic, rather than the result of damage and contamination?	Mastodon and *T. rex*	No	Errors in interpretation of spectra [see table S1 and (*13*)]? Damage-induced errors in sequence
Do the results make sense, and are there enough data to make the study useful and/or to support the conclusions?	Mastodon	Yes	Weak affinity to mammals
	T. rex	No	Affinity of α1(I) peptides to amphibians, not birds or reptiles

[1] BioArch, Departments of Biology, Archaeology, Chemistry and Technology Facility, University of York, Post Office Box 373, York YO10 5YW, UK. [2] Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI 48109–0404, USA. [3] Evolutionary Biology Group, Department of Zoology, University of Oxford, OX1 3PS, UK. [4] Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103, Leipzig, Germany. [5] Department of Zoology, Michigan State University, East Lansing, MI 48824, USA. [6] Department of Biology, Pennsylvania State University, University Park, PA 16802, USA. [7] Center for Biomolecular Science and Engineering, University of California–Santa Cruz, CA 95064, USA. [8] Department of Parasitology, Kuvin Center, Hebrew University of Jerusalem, Israel. [9] Biological Institute, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark. [10] Chemistry Department, Trinity College, 300 Summit Street, Hartford, CT 06106, USA. [11] Cleveland Museum of Natural History, 1 Wade Oval Drive, University Circle, Cleveland, OH 44106, USA. [12] National Museum of Natural History "Naturalis," P.O. Box 9517, 2300 RA Leiden, Netherlands. [13] National Heritage Trust Fund Mauritius, Mauritius Institute, La Chaussée Street Port Louis, Mauritius.

*To whom correspondence should be addressed. E-mail: mc80@york.ac.uk

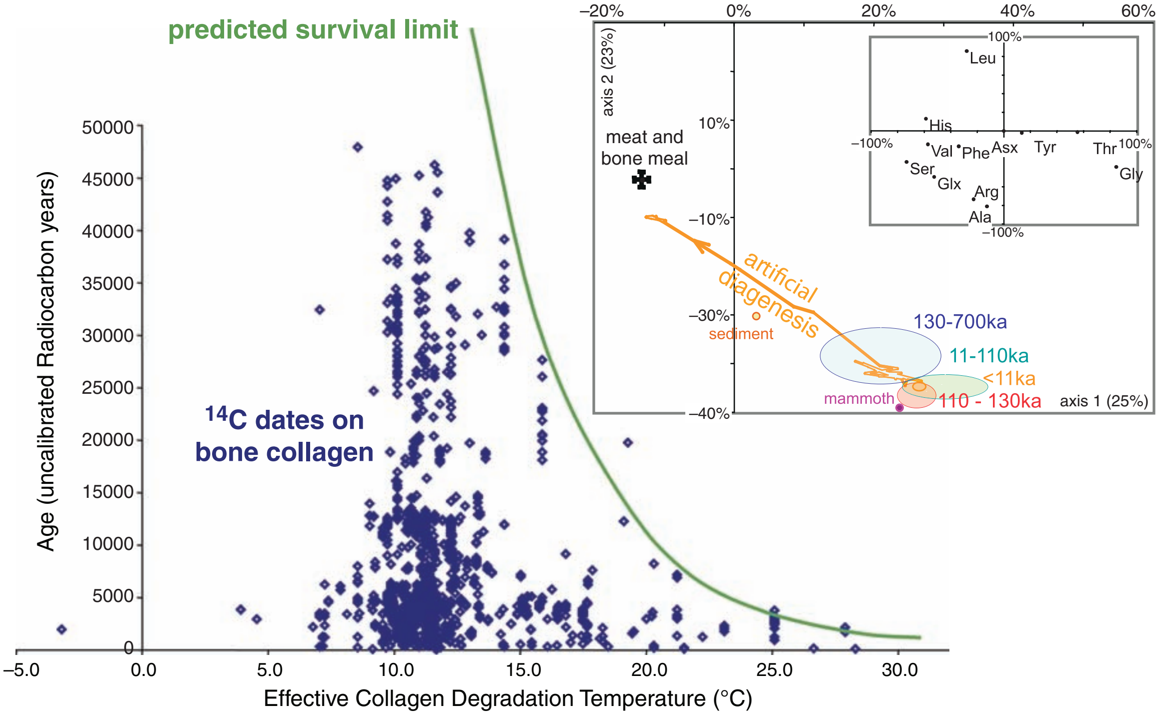


Fig. 1. Plot of radiocarbon age versus estimated effective collagen degradation temperature for radiocarbon-dated bones from laboratory databases (principally Oxford and Groningen). The line represents the expected calendar age at which 1% of the original collagen remains following a zero-order reaction; almost no bone collagen survives beyond this predicted limit. (**Inset**) The 99% confidence intervals of amino acid compositions by first two principal component analyses (48% of total variance) for bones from NW Europe aged <11 ky (*n* = 324), 11 to 110 ky (*n* = 210), 110 to 130 ky (*n* = 26), and 130 to 700 ky (*n* = 31). Pliocene samples are not plotted, as their composition (*n* = 8) is highly variable and yields of amino acids are low. The orange line indicates a compositional trend observed when compact bone is heated for 32 days at 95°C, which reduces collagen to 1% of the initial concentration [each inflection represents a separate analysis; *n* = 32)]. The composition becomes more similar to mixed tissue samples (meat and bone meal; *n* = 32), principally due to the depletion of Gly. An amino acid profile for mammoth is consistent with collagen, unlike the associated sediment sample [data from (*11*)].

authors, obvious contamination sources such as animal glue (used in conservation) can be excluded. However, concentrating protein from the large amounts of bone used (2.5 g) may have heightened the risk of extraneous proteins entering the sample during extraction, although there have been no systematic studies of this phenomenon. Independent extraction and analyses would have strengthened claims for the authenticity of the origin of the peptides (and potentially ameliorated the original problems of data interpretation) (*4*).

The remarkable soft-tissue preservation of the investigated *T. rex* specimen (MOR 1125) has been documented (*8*). However, microscopic preservation does not equate with molecular preservation (*9*). Immunohistochemistry provides support for collagen preservation, but Asara *et al.* (*2*) presented no data regarding inhibition assays with collagen from different species or cross-reactivity with likely contaminants [e.g., fungi (*10*)]. Curiously, no amino acid compositional analysis was conducted [see (*11*)], although immonium ions were identified by time-of-flight secondary ion mass spectrometry. In our experience, collagen-like amino acid profiles have been obtained in all bones from which we could obtain collagen sequence (Fig. 1, inset).

Regarding the proof of sequence authenticity, the spectra reported by Asara *et al.* (*12*) are inconsistent with some of the sequence assignments (*13*) (table S1). A common diagenetic modification, deamidation, not considered in (*2*), may shed light on authenticity. The facile succinimide-mediated deamidation (*14*) of asparagine occurred at $N_{229}G$ and $N_{1156}G$ in ostrich peptides (Ost 4 and Ost5) (see table S1 for nomenclature), presumably during sample preparation. Direct hydrolytic deamidation is slower (*14*), and an expectation of elevated levels of such products is reasonable for old samples. We agree with the most recent interpretation (*13*) of the spectrum illustrated in Fig. 2B as α1(I) $G_{362}SEGPEGVR_{370}$, the deamidated ($Q \rightarrow E_{367}$) form of the sequence found in most mammals (*12*). By way of contrast, none of the three glutamine residues in the reported *T. rex* peptides are deamidated (table S1). Only time will tell if Q→E is a useful marker for authentically old collagen, but from the evidence presented, the mastodon sequence looks more diagenetically altered than *T. rex*.

The unusual, fragmented nature of the reported *T. rex* sequence does not make it amenable to standard, model-based phylogenetic analysis. Instead, we examined the phylogenetic signal of the α1(I) fragments of mastodon and *T. rex* using Neighbor-Net analysis and uncorrected genetic distances. Using the sequences reported in (*13*), both the *T. rex* and mastodon signal display an affinity with amphibians (Fig. 2A). Our reinterpretation of the spectra (*12*) changes the affinity of mastodon but not of *T. rex* (Fig. 2B). In addition to the α1(I) peptides used in the Neighbor-Net analysis, Asara *et al.* reported two other peptides from *T. rex* (*13*); we question the interpretation of the α1(II) spectra (identical to frog) but not the α2(I) spectra (identical to chicken).

We require more data to be convinced of the authenticity of the *T. rex* collagen sequences reported by Asara *et al.* Nevertheless, the handful of spectra reported for the temperate Pleistocene mastodon fail neither phylogenetic nor diagenetic tests, thus highlighting the potential of protein mass spectrometry to bridge the present gulf in our understanding between the fate of archaeological and fossil proteins. To avoid past mistakes of ancient DNA research (*1*), we recommend that future fossil protein claims be considered in light of tests for authenticity such as those presented here.

Reference and Notes

1. E. Willerslev, A. Cooper, *Proc. R. Soc. London. B. Biol. Sci.* **272**, 3 (2005).
2. J. M. Asara, M. H. Schweitzer, L. M. Freimark, M. Phillips, L. C. Cantley, *Science* **316**, 280 (2007).
3. M. H. Schweitzer, *Palaeontologia Electronica* **5**, editorial 2 (2003); http://palaeo-electronica.org/2002_2/r_and_p.pdf
4. M. T. P. Gilbert, H.-J. Bandelt, M. Hofreiter, I. Barnes, *Trends Ecol. Evol.* **20**, 541 (2005).
5. M. J. Collins, M. S. Riley, A. M. Child, G. Turner-Walker, *J. Archaeol. Sci.* **22**, 175 (1995).
6. C. A. Miles, M. Ghelashvili, *Biophys. J.* **76**, 3243 (1999).
7. K. R. Johnson, D. J. Nichols, J. H. Hartman, *The Hell Creek Formation and the Cretaceous-Tertiary Boundary in the Northern Great Plains: Geological Society of America Special Paper* **361**, 503–510 (2002).
8. M. H. Schweitzer *et al.*, *Science* **316**, 277 (2007).
9. N. S. Gupta, D. E. G. Briggs, R. D. Pancost, *J. Geol. Soc. London* **163**, 897 (2006).
10. P. Sepulveda *et al.*, *Infect. Immun.* **63**, 2173 (1995).

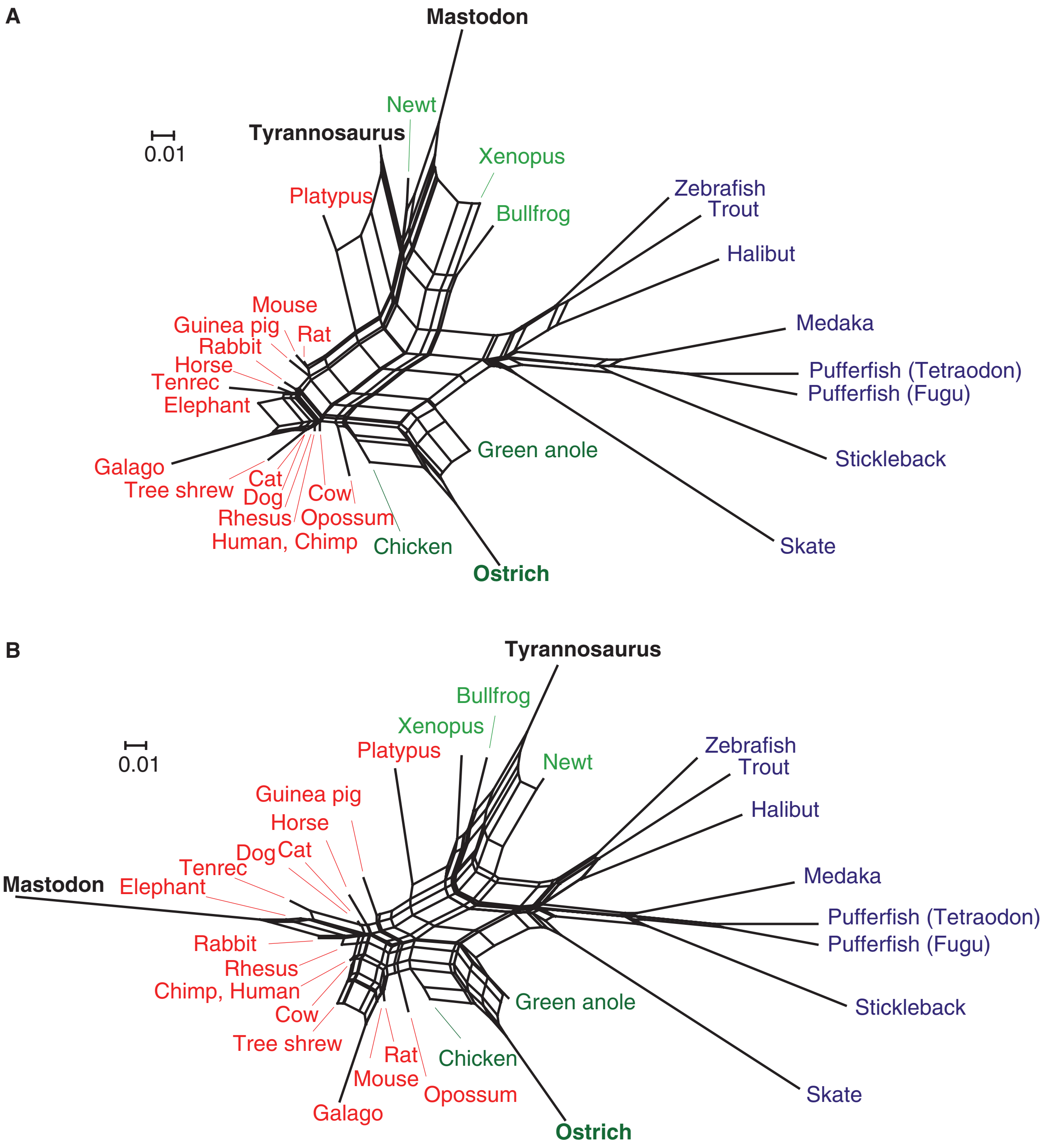


Fig. 2. Phylogenetic networks of α1(I) sequences using Neighbor-Net analysis (**A**) with the most recent Asara *et al*. assignments (*13*) and (**B**) after our reinterpretation of the mass spectrometric data (*12*). *T. rex* does not group with bird/reptile using either set of sequence alignments. More sequence is required for a full, model-based phylogenetic analysis.

11. M. Schweitzer, C. L. Hill, J. M. Asara, W. S. Lane, S. H. Pincus, *J. Mol. Evol.* **55**, 696 (2002).
12. See supporting material on *Science* Online.
13. J. M. Asara *et al.*, *Science* **317**, 1324 (2007).
14. N. E. Robinson *et al.*, *J. Pept. Res.* **63**, 426 (2004).
15. This work was supported by NSF (EAR-0309467), National Environment Research Council (NE511148, GR9/01656), National Center for Research Resources (P41-18627), European Commission (MEST-CT-2004-007909, MEST-CT-2005-020601), Wellcome Trust Bioarchaeology Fellowships (K.P. and H.K.) Analytical Chemistry Trust Fund, the Royal Society of Chemistry Analytical Division, Engineering and Physical Sciences Research Council, and the Michigan Proteome Consortium.

Supporting Online Material
www.sciencemag.org/cgi/content/full/319/5859/33c/DC1
SOM Text
Table S1
References

26 June 2007; accepted 20 November 2007
10.1126/science.1147046

5.7.2 Publication II

The following paper was published on *Nature Chemistry* in 2010:

Hermann Ehrlich, Rainer Deutzmann, Eike Brunner, Enrico Cappellini, Hannah Koon,Caroline Solazzo, **Yue Yang**, David Ashford, Jane Thomas-Oates, Markus Lubeck,Carsten Baessmann, Tobias Langrock, Ralf Hoffmann, Gert Wo¨rheide, Joachim Reitner,Paul Simon, Mikhail Tsurkan, Aleksandr V. Ereskovsky, Denis Kurek, Vasily V. Bazhenov,Sebastian Hunoldt, Michael Mertig, Denis V. Vyalikh, Serguei L. Molodtsov, Kurt Kummer,Hartmut Worch, Victor Smetacek and Matthew J. Collins "Mineralization of the meter-long biosilica structures of glass sponges is templated on hydroxylated collagen", *Nature Chemistry*, 2(2010) 1084–1088

In this paper, UniColl, which was referred to the in-house 'Collagens' database, was used as an assistant tool to sequence collagen from the metre-long stalk of the glass rope sponge (Hyalonema sieboldi; Porifera, Class Hexactinellida). By pattern recognition, mass spectra with peptide sequences containing the G-X-Y collagen motif were selected for manual de novo sequencing. The new sequences obtained with this approach were uploaded onto UniColl. While searching samples against the supplemented UniColl, a hydroxylated fibrillar collagen that contains an unusual [Gly-3Hyp-4Hyp] motif was revealed to compose the organic fraction, and possibly to be predisposed for silica precipitation and provide a novel template for biosilicification in nature.

The full text is supplemented as follows.

ARTICLES
PUBLISHED ONLINE: 23 NOVEMBER 2010 | DOI: 10.1038/NCHEM.899

nature chemistry

Mineralization of the metre-long biosilica structures of glass sponges is templated on hydroxylated collagen

Hermann Ehrlich[1]*, Rainer Deutzmann[2], Eike Brunner[1], Enrico Cappellini[3], Hannah Koon[3], Caroline Solazzo[3], Yue Yang[3], David Ashford[4,5], Jane Thomas-Oates[5,6], Markus Lubeck[7], Carsten Baessmann[7], Tobias Langrock[8], Ralf Hoffmann[8], Gert Wörheide[9], Joachim Reitner[10], Paul Simon[11], Mikhail Tsurkan[12], Aleksander V. Ereskovsky[13,14], Denis Kurek[1,15], Vasily V. Bazhenov[1,16], Sebastian Hunoldt[1], Michael Mertig[17], Denis V. Vyalikh[18,19], Serguei L. Molodtsov[18,19], Kurt Kummer[18,19], Hartmut Worch[17], Victor Smetacek[20] and Matthew J. Collins[3]*

The minerals involved in the formation of metazoan skeletons principally comprise glassy silica, calcium phosphate or carbonate. Because of their ancient heritage, glass sponges (Hexactinellida) may shed light on fundamental questions such as molecular evolution, the unique chemistry and formation of the first skeletal silica-based structures, and the origin of multicellular animals. We have studied anchoring spicules from the metre-long stalk of the glass rope sponge (*Hyalonema sieboldi*; Porifera, Class Hexactinellida), which are remarkable for their size, durability, flexibility and optical properties. Using slow-alkali etching of biosilica, we isolated the organic fraction, which was revealed to be dominated by a hydroxylated fibrillar collagen that contains an unusual [Gly–3Hyp–4Hyp] motif. We speculate that this motif is predisposed for silica precipitation, and provides a novel template for biosilicification in nature.

Among the different biominerals, silica in its different amorphous forms is probably the most intriguing. It is probably the first and oldest natural bioskeleton, with unique mechanical properties and an extremely high specific surface area. Of the challenging topics that are receiving renewed attention today, the study of the mechanisms of biosilicification including the specificity of organic templates is among the most fascinating from chemical, biological and materials points of view.

Although it was first proposed by the groups of Morse[1,2] and Müller[3] that low molecular weight proteins—silicateins—play a pivotal role for the silification of spicules in the sponge class Demospongiae, the situation in the other siliceous spicule-producing sponge class, Hexactinellida, is less clear.

Hexactinellids are phylogenetically among the oldest metazoans, established in the Late Protoerozoic. Their skeleton is composed of silica-based spicules, the largest of which project from the body surface and serve as protective lateral spines or basal attachment roots. The basal twisted column of root tuft spicules in the 'glass rope sponge' (*Hyalonema sieboldi*; Porifera, Class Hexactinellida; Fig. 1a,b) can extend up to one metre in length and acts by anchoring the sponge in the soft bottom sediment. These spicules, which have remarkable optical properties[4], are both durable (deep-sea glass sponges live for centuries)[5] and flexible (they can be bent full circle)[6].

The presence of silicatein has been reported in the non-anchoring body microspicules of the hexactinellid *Crateromorpha meyeri*[7]. The metre-long anchoring spicules of *Hyalonema sieboldi* have been shown to be hierarchically structured[4,6], but the nature of the organic template on which silica is deposited has eluded identification. Using a novel, slow-etching method[8], we have previously reported collagen-like fibrillar proteins within both *H. sieboldi*[6,8] and *Monorhaphis chuni*[9,10] glass sponges. In this Article, we report the first detailed characterization of this.

Results and discussion

The C 1*s* near-edge X-ray absorption fine structure spectrum of *H. sieboldi* spicules shares characteristic features with a vertebrate collagen standard (Supplementary Fig. S1), as does ^{13}C solid-state nuclear magnetic resonance (NMR) spectroscopy of isolated fibrillar protein (Supplementary Fig. S2). In contrast to the identification of collagen in spicules of the glass sponge *Euplectella* sp.[11], which is probably a contaminant, we have isolated between 250 and 300 mg fibrillar protein per gram of glassy spicule from *H. sieboldi* (Fig. 1d). Polyclonal antibodies detected type I (but not type IV) collagen in the root spicules of *Hyalonema* sp. (Supplementary Fig. S3), but neither type I nor type IV collagens were detected in spicules of the demosponge *Petrosia* sp. or in the spicules supporting the body of two other hexactinellid species of the family Rosselidae. These results are consistent with an alternative macromolecular template to collagen for silicification, for example, chitin in the glass sponge *Rossella fibulata*[12].

A short digestion with papain of material solubilized by overnight digestion with trypsin, were purified by reversed-phase chromatography and subject to Edman degradation. All peptides that were sequenced had the characteristic [Gly–Xaa–Yaa]$_n$ repeat. 3-Hyp and 4-Hyp were found exclusively in positions Xaa and Yaa, respectively, and were identified unambiguously by comparison with authentic PTH amino-acid standards (Box 1).

The extracted collagen (Fig. 2a) was analysed by mass spectrometry using three different approaches (Supplementary Sections S8–S9d). To confidently assign the sequences and to differentiate

A full list of affiliations appears at the end of the paper.

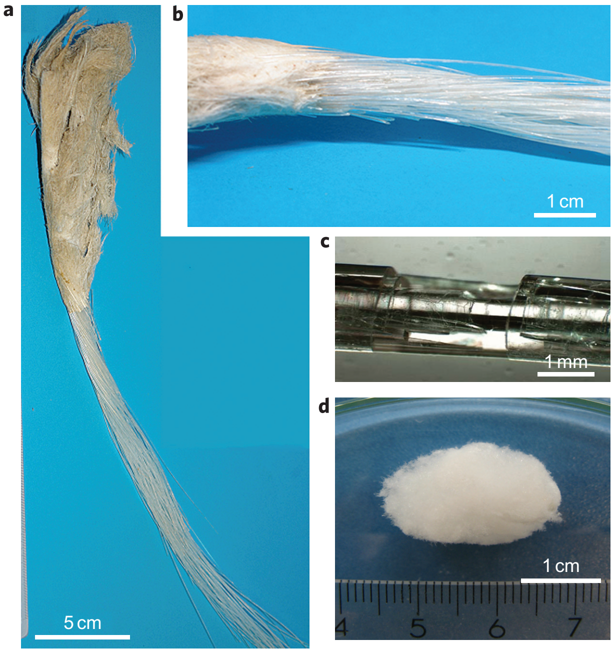


Figure 1 | Marine glass sponge *Hyalonema sieboldi*, a typical member of the Hyalonematidae family. a, Image of marine glass sponge *Hyalonema sieboldi.* Anchoring spicules of these sponges (**b**) have a multilayered structure and are organized according to the principle of 'cylinder in cylinder' (**c**). **d**, Fibrillar protein of a collagenous nature was isolated from the spicules using gentle desilicification in alkaline solution.

hydroxyproline (residue mass 113.04768 Da) from Leu/Ile (residue mass 113.08406 Da), a high-resolution, high mass accuracy mass spectrometer was used to carry out liquid chromatography-electrospray-tandem mass spectrometry (LC-ESI-MS/MS) analyses of the tryptic digest of the extracted collagen sample. The resulting data were searched against public and in-house protein sequence databases, and a range of peptides was identified as originating from collagen (Table 1). The spectra were also independently interpreted *de novo*, which resulted in the same assignments (Table 1).

Evidence for the peptide GAQGPLGPR identified from Edman sequencing (Box 1) was also obtained by mass spectrometry (Table 1), all other MS peptides were novel and most display the characteristic [Gly–Xaa–Yaa]$_n$ motif. For the longest peptide, the complete y_7–y_{13} ion series supports the existence of an unusual double hydroxylation 'Gly–Hyp–Hyp' (Supplementary Fig. S5). Spectra were also obtained that were confidently attributed to cytoskeletal actin (accession no. 3386376 from the ascidian *Molgula oculata*; Supplementary Table S1). Glass sponges lack contractile tissues, but thick actin microfilament bundles extending for hundreds of micrometres have been reported to form the core of the blunt giant rod-like extensions projecting from the edges of syncytial aggregates[13]. Actin microfilaments have been previously observed to be associated with silica-deposition vesicles in protists (Synurophyceae) and diatoms, where they are thought to be involved in shaping the cytoplasm. Their association with the organic spicule, entombed in silica, suggests a different role, perhaps associated with the maturation of the silica fibre.

Box 1 | Edman degradation

GAQ G(P)L GP
IGPDEPLKGKI
GGF GLO GR
G(V)D GNO GIX GAT GS
GGS GAO GLO GAI GNQ GAO
V GDO GLV GDL GAQ GPQ GSQ GLV G
GIO GPQ GFT GAI GVT GSO GEI GAO G
GSV **GOO** GNO GVQ GVS GO
GAT **GOO** GIS **GOO** GPQ GQO GTO GI
I GPA GPQ GQO **GOO** GPG GPX **GOO**

Amino-acid analysis of the collagen isolated from the *H. sieboldi* spicule showed a Pro/Hyp ratio of 1.33, and a ratio of ~3:1 of trans-4-Hyp to trans-3-Hyp (Fig. 2c; Supplementary Table S2). These ratios are remarkably consistent with peptide sequence data, with hydroxylation of 33% of those Pro residues in the Xaa position and 100% in the Yaa position (as 3-Hyp and 4-Hyp, respectively) of the [Gly–Xaa–Yaa]$_n$ motif. This result is similar to those reported previously for *Geodia cydonium* sponge collagen[14]. The presence of both 4-hydroxyproline and 3,4-dihydroxyproline has been reported in siliceous cell walls of diatoms[15,16], and these authors suggested that hydroxylated amino acids could play a role in silicification of diatom cell walls. Hydroxy amino acids are known to be distributed in cell walls of diatoms[17] as well as in silicateins, the specific proteins responsible for silicification in demosponges[1–3].

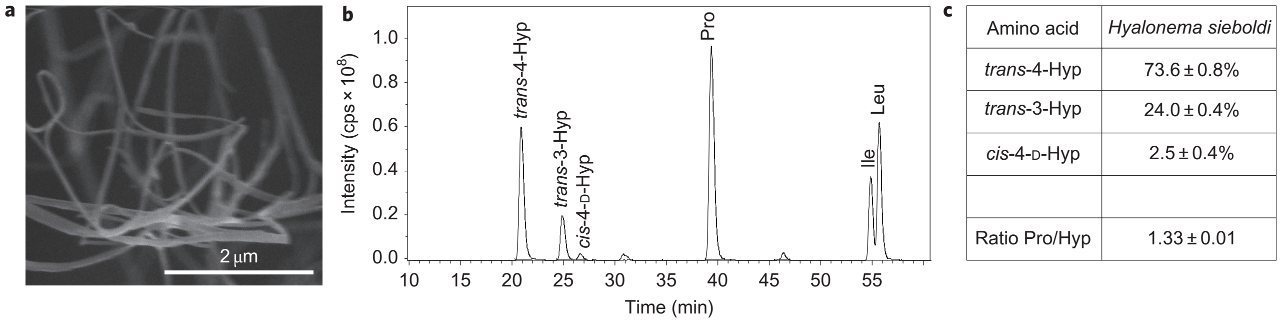
We were able to demonstrate the role of the hydroxylation state of collagen in silica polycondensation. The rate of silica formation was significantly higher in *H. sieboldi* spicular collagen than it was in two samples of collagen lacking significant trans-3-Hyp that were isolated from calf skin and *Chondrosia reniformis* (a non-spicular desmosponge) mesohyl (Supplementary Fig. S15). However, we were able to reduce silicification activity when the 3- and 4-hydroxyproline residues of *H. sieboldi* were protected by formation of a ketal group (Fig. 3). Functional recovery was restored when the ketal protecting group was removed (Fig. 3d, inset).

The collagen motif determined in *Hyalonema* is consistent with the model of Schumacher and colleagues[18], which describes 3(S)-hydroxyproline residues in the Xaa position of the collagen triple helix. This structure offers a plausible molecular model for the interaction between polysilicic acid and Gly–3Hyp–4Hyp polypeptides of isolated glass sponge collagen (Supplementary Fig. S12). It is established that the interaction between orthosilicic acid and hydroxyl groups is likely to be a hydrogen bond[19]. Our model shows the possibility of stable complex formation on the basis of hydrogen bonding between hydroxyl groups of polysilicic acid and surface exposed hydroxyls of 3-Hyp and 4-Hyp. Our model proposes a functional role for trans-3-Hyp in sponge collagen silicification. Collagen will

Table 1 | Collagen peptides identified by high-resolution mass spectrometry and manual *de novo* sequencing, or by Edman sequencing.

MS/MS	Observed	M_r (expt.)	M_r (calc.)	ppm	Score	Expect
GPJ GPT GJQ GAR*	562.8033	1,123.592	1,123.5986	−5.87	77	0.000015
E GEJ GJO GET GPR	664.3041	1,326.5937	1,326.6052	−8.64	16	0.024
GJO GAO GJD GNO GPA GJR	832.9138	1,663.8130	1,663.8166	−2.13	86	8.9×10^{-9}
(MEGPT) GAP GAO GDA GVJ **GOO GOO** G(PQGPR)	896.4121	2,686.2146	2,686.2294	−5.51	66	1.1×10^{-6}
GAQ GPJ GPR	426.7374	851.4603	851.4613	−1.19	54	6.7×10^{-6}

Q/N, glutamine/asparagine deamidation; O, hydroxyproline; J indicates leucine or isoleucine; the presence of brackets indicates uncertain residue (P) or sequence order (for example, (XY) indicates either XY or YX). *A significant match to the starlet sea anemone *Nematostella vectensis* (phylum *Cnidaria*; gi|156394292).



Amino acid	*Hyalonema sieboldi*
trans-4-Hyp	73.6 ± 0.8%
trans-3-Hyp	24.0 ± 0.4%
cis-4-D-Hyp	2.5 ± 0.4%
Ratio Pro/Hyp	1.33 ± 0.01

Figure 2 | Analysis of the isolated spicular organic matrix. a, SEM image of the nanofibrils observed in alkali extracts obtained after gentle demineralization over 14 days at 37 °C. **b**, Extracted ion chromatogram for the N_2-(5-fluoro-2,4-dinitrophenyl)-L-valine amide (FDVA) derivatives at m/z 412 (Hyp, Leu, Ile) and m/z 396 (Pro) from the hydrolysate of the organic matrix of demineralized *H. sieboldi* spiculae. cps, counts per second. **c**, Relative amounts of all three Hyp-isomers and the Pro/Hyp ratio based on the signal intensities shown in **b**. The table summarizes the data for the amino-acid analysis, which provides the amount of each amino acid (g or mol). As all amino acids represent isomers with identical mass, the percentage represents the content of each Hyp residue in collagen.

present a layer of hydroxyl groups that can undergo condensation reactions with silicic acid molecules with a consequent loss of water. As a result, the initial layer of condensed silicic acid will be held fixed to the collagenous template in a geometric arrangement that will favour further polymerization of silicic acid, similar to the model proposed by Hecky and colleagues[16]. It therefore appears that collagen was a novel template for biosilicification that emerged at an early stage during metazoan evolution, and that the occurrence of additional trans-3-Hyp plays a key role in stabilizing silicic acid molecules and initiating the precipitation of silica.

Hydroxylated collagen appears to form the basis for the extraordinary mechanical and optical properties of hexactinellid spicules[20]. The self-assembly properties of collagen and its templating activity with respect to silicification are consistent with recent ideas on the development of hierarchical silica-based architectures[21]. Macroscopic bundles of silica nanostructures result from the kinetic cross-coupling of two molecular processes: a dynamic supramolecular self-assembly and a stabilizing silica mineralization. The feedback interactions between template growth and inorganic deposition are driven non-enzymatically by means of hydrogen bonding. We speculate that the hydroxylated glass sponge collagen may change the nature of silica in aqueous solution by converting the distribution of oligomers into a more uniform and useful set of nanoparticle precursors for assembly into the growing solid (Supplementary Section S10).

Our findings suggest that in addition to the previously described silicatein-based biosilicification of sponge spicules, collagen has a key role to play in the formation of the long, flexible, optically pure anchoring spicules of the Hexactinellids. Increased atmospheric oxygen during the Proterozoic may have been linked to post-translational hydroxylation of proline and lysine residues[22], and it is tempting to speculate that the occurrence of silica- and hydroxylated collagen-based composites in skeletal structures of the first metazoan might be a co-evolutionary event. A reconstruction of the evolution of biocalcification as well as of biosilicification with respect to collagen may be a key way to obtain strong evidence

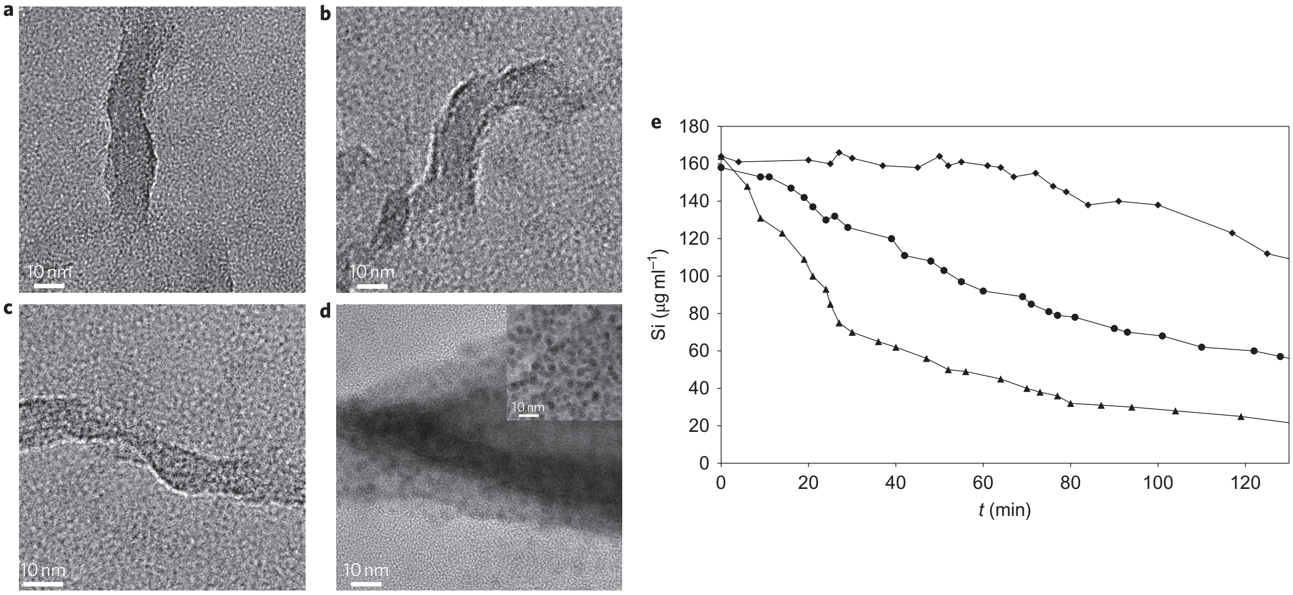


Figure 3 | HR-TEM images of silicification on *H. sieboldii* collagen. Silicification is apparent as nanoparticles after exposure of nanofibrillar *H. sieboldii* spicular collagen (**a**) to a solution of sodium methasilicate solution for 30 min. **b**, However, after protection of 3- and 4-hydroxyproline residues by ketal groups (Supplementary Figs S11 and S13), there is no visible silica deposition. **c**, Cleavage of the ketal protecting groups from collagen leads to a functional recovery with respect to silicification. **d**, The layer of silica nanoparticles is formed around the nanofibril of native spicular collagen during the first 30 min of silicification, as seen in the native collagen fibre (Supplementary Fig. S16). **e**, The results are in good agreement with measurements of activity (Supplementary Fig. S10) for non-protected collagen (filled triangle), which is lost following protection (filled diamonds), and partially restored when this protection is removed (filled circles).

of ancient, ancestral programs[23] of biomineralization based on this common template. The bioconstruction of the uniquely large siliceous structures (ten orders of magnitude longer than the spicules of demosponges) was probably enabled by the incorporation of collagen, which can play a role as a template and also provide structural support. This may mean a re-thinking of the role of collagen in the evolution of biomineralization, and almost certainly opens up new strategies for the biomimetic synthesis of silica-based materials.

Methods

Basal spicules of *Hyalonema sieboldi* (Hexactinellida: Porifera) Gray 1835, collected from a depth of 5,000 m in the Pacific (12 °N, 137 °E), as well as those from *H. sieboldi* collected by C. Eckert in Sagami Bay, Japan (FS 'Tansai Maru', St. TS 4–8, May 2004) were used in this study. Dried spicules were washed three times in distilled water, cut into pieces (2–5 cm long), and placed in a solution containing purified *Clostridium histolyticum* collagenase (Sigma) to digest any possible collagen contamination of an exogenous nature. After incubation for 24 h at 15 °C, the pieces of spicule were washed again, three times in distilled water, then dried and placed in 10 ml plastic vessels containing 5 ml of 2.5 M NaOH (Fluka) solution. The vessel was covered, placed under thermostatic conditions at 37 °C and shaken slowly for 14 days. Alkali extracts of *H. sieboldi* spicules containing fibrillar protein were dialysed against deionized water on Roth (Germany) membranes with a cut-off of 14 kDa. The dialysed material was dried under vacuum in a CHRIST lyophilizer (Germany) and used for collagen identification (described in detail in the Supplementary Information).

Analytical methods. The analytical methods used in this work include near-edge X-ray absorption fine structure spectroscopic method (NEXAFS), ^{13}C solid-state nuclear magnetic resonance (NMR), scanning electron microscopy (SEM), transmission electron microscopy (TEM), high-resolution transmission electron microscopy (HR-TEM), Fourier transform infrared spectroscopy (FTIR) and mass spectrometric methods, Edman degradation, the immunochemical method, ketal protection of the hydroxyl groups of the 3-hydroxyproline and 4-hydroxyproline of collagen, as well as measurements of the silica condensing activity of collagens). These are described in detail in the Supplementary Information.

Data deposition. All the raw mass spectrometric data associated with this manuscript may be downloaded from ProteomeCommons.org Tranche, http://tranche.proteomecommons.org, using the hash

VmBiB25HJz + NtxKBunAPvmvsR0Ax8ddFDBVdziCJKC9i4
gLDSQeYQoqO/NP7RrTf89 RfZSMilHLn/OA0aWhn +
Nyy8KIAAAAAAA4xYg = =

or the link

http://www.proteomecommons.org/data-downloader.jsp?fileName =
VmBiB25HJz%2BNtxKBunAPvmvsR0Ax8ddFDBVdziCJKC9i4gLDSQeYQoqO/
NP7RrTf89RfZSMilHLn/OA0aWhn%2BNyy8KIAAAAAAA4xYg = =

The hash may be used to demonstrate exactly which files were published as part of the data set of this manuscript, and also to check that the data have not changed since publication.

Received 2 February 2010; accepted 5 October 2010;
published online 23 November 2010

References

1. Shimizu, K., Cha, J., Stucky, G. D. & Morse, D. E. Silicatein alpha: cathepsin L-like protein in sponge biosilica. *Proc. Natl Acad. Sci. USA* **95,** 6234–6238 (1998).
2. Cha, J. N. *et al.* Silicatein filaments and subunits from a marine sponge direct the polymerization of silica and silicones *in vitro*. *Proc. Natl Acad. Sci. USA* **96,** 361–365 (1999).
3. Müller, W. E. G. *et al.* Silicateins, the major biosilica forming enzymes present in demosponges: protein analysis and phylogenetic relationship. *Gene* **395,** 62–71 (2007).
4. Müller, W. E. G. *et al.* Unique transmission properties of the stalk spicules from the hexactinellid *Hyalonema sieboldi*. *Biosens. Bioelectron.* **21,** 1149–1155 (2006).
5. Dayton, P. K. Observations of growth, dispersal and population dynamics of some sponges in McMurdo Sound, Antarctica, in *Colloques internationaux du C.N.R.S. 291. Biologie des spongiaires* (eds Levi, C. & Boury-Esnault, N.), 271–282 (Editions du Centre National de la Recherche Scientifique, 1979).
6. Ehrlich, H. & Worch, H. Collagen, a huge matrix in glass-sponge flexible spicules of the meter-long *Hyalonema sieboldi*, in *Handbook of Biomineralization Vol. 1. The Biology of Biominerals Structure Formation* (ed. Bäuerlein, E.) 23–41 (Wiley VCH, 2007).
7. Müller, W. E. G. *et al.* Silicatein expression in the hexactinellid *Crateromorpha meyeri*: the lead marker gene restricted to siliceous sponges. *Cell Tissue Res.* **333,** 339–351 (2008).
8. Ehrlich, H. *et al.* A modern approach to demineralization of spicules in glass sponges (Porifera: Hexactinellida) for the purpose of extraction and examination of the protein matrix. *Russ. J. Marine Biol.* **32,** 186–193 (2006).
9. Ehrlich, H. *et al.* Nanostructural organization of naturally occurring composites. Part I. Silica-collagen-based biocomposites. *J. Nanomater.* doi:10.1155/2008/623838 (2008).
10. Ehrlich, H., Heinemann, S., Hanke, T. & Worch, H. Hybrid materials from a silicate-treated collagen matrix, methods for the production thereof and the use thereof. International patent WO2008/023025 (2008).
11. Travis, D. F., Francois, C. J., Bonar, L. C. & Glimcher, M. J. Comparative studies of the organic matrices of invertebrate mineralized tissues. *J. Ultrastruct. Res.* **18,** 519–550 (1967).
12. Ehrlich, H. *et al.* Nanostructural organization of naturally occurring composites. Part II. Silica-chitin-based biocomposites. *J. Nanomater.* doi:10.1155/2008/670235 (2008).
13. Leys, S. P. Cytoskeletal architecture and organelle transport in giant syncytia formed by fusion of hexactinellid sponge tissues. *Biol. Bull.* **188,** 241–254 (1995).
14. Diehl-Seifert, B. *et al.* Attachment of sponge cells to collagen substrata: effect of a collagen assembly factor. *J. Cell Sci.* **79,** 271–285 (1985).
15. Nakajima, T. & Volcani, B. E. 3,4-Dihydroxyproline: a new amino acid in diatom cell walls. *Science* **164,** 1400–1401 (1969).
16. Hecky, R. E., Mopper, K., Kilham, P. & Degens, E. T. The amino acid and sugar composition of diatom cell walls. *Mar. Biol.* **19,** 323–331 (1973).
17. Sadava, D. & Volcani, B. E. Studies on the biochemistry and fine structure of silica shell formation in diatoms. Formation of hydroxyproline and dihydroxyproliner in *Nitzschia angularis*. *Planta* **135,** 7–11 (1977).
18. Schumacher, M. A., Mizuno, K. & Bachinger, H. P. The crystal structure of a collagen-like polypeptide with 3(S)-hydroxyproline residues in the Xaa position forms a standard 7/2 collagen triple helix. *J. Biol. Chem.* **281,** 27566–27574 (2006).
19. Tilburey, G. E., Patwardhan, S. V., Huang, J., Kaplan, D. L. & Perry, C. C. Are hydroxyl-containing biomolecules important in biosilicification? A model study. *J. Phys. Chem. B* **111,** 4630–4638 (2007).
20. Kulchin, Y. N. *et al.* Optical fibres based on natural biological minerals—sea sponge spicules. *Quantum Electron.* **38,** 51–55 (2008).
21. Pouget, E. *et al.* Hierarchical architectures by synergy between dynamical template self-assembly and biomineralization. *Nature Mater.* **6,** 434–439 (2007).
22. Exposito, J.-Y., Cluzel, C., Garrone, R. & Lethias, C. Evolution of collagens. *Anat. Rec.* **268,** 302–316 (2002).
23. Livingston, B. T. *et al.* A genome-wide analysis of biomineralization-related proteins in the sea urchin *Strongylocentrotus purparatus*. *Dev. Biol.* **300,** 335–348 (2006).

Acknowledgements

This work was partially supported by NE/C511148/1 grant to M.J.C. and J.T.-O., by a joint grant from Deutscher Akademischer Austauschdienst (DAAD) (grant ref. 325; A/08/72558) and Russian Ministry of Education and Science (RMES) (AVCP grant no. 8066) to V.V.B., and by the Erasmus Mundus External Co-operation Programme of the European Union 2009 to D.K. The Centre of Excellence in Mass Spectrometry (CoEMS) at York is supported by Science City York and Yorkshire Forward using funds from the Northern Way Initiative. The authors also thank H. Lichte for use of the facilities at the Special Electron Microscopy Laboratory for high-resolution and holography at Triebenberg, TU Dresden, Germany. The authors thank G. Bavestrello, K. Tabachnick, M. Hofreiter, H. Roempler and E. Bäeuerlein for their helpful discussions. Finally, the authors are very grateful to S. Paasch, H. Meissner, G. Richter, T. Hanke, O. Trommer, A. Mensch, S. Heinemann for excellent technical assistance. G.W. acknowledges funding from the German Science Foundation (DFG). Financial support by DFG (Graduiertenkolleg 378) to R.H. and T.L. and the European Fund for Regional Structure Development (EFRE, European Union and Free State Saxonia) to R.H. is gratefully acknowledged.

Author contributions

All authors contributed to the design or execution of experiments, or analysed data. H.E. supervised the experiments, carried out demineralization experiments, performed collagen isolation, and wrote the manuscript. P.S. performed SEM and HRTEM, and prepared figures. A.E. collected, prepared and identified sponge samples and contributed to writing the manuscript. M.M., D.V.V., K.K. and S.L.M. performed NEXAFS experiments and designed figures. M.T. and V.V.B. carried out collagen modification. S.H. performed FTIR and prepared figures. E.B. performed NMR. R.D. performed Edman degradation and R.H. and T.L. performed amino acid analysis and mass spectrometry. M.C., H.K., C.S., Y.Y., E.C., D.A., M.L., C.B. and J.T.-O. were involved in acquiring and interpreting the mass spectrometric data, and M.C., H.K., E.C., D.A. and J.T.-O contributed to the writing of the manuscript. H.W., M.C., H.E., G.W., J.R., V.S. and E.B. analysed the results with regard to evolutionary implications and mechanisms of biomineralization, designed concepts, and wrote the manuscript.

Additional information

The authors declare no competing financial interests. Supplementary information accompanies this paper at www.nature.com/naturechemistry. Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/. Correspondence and requests for materials should be addressed to H.E. and M.J.C.

[1]Institute of Bioanalytical Chemistry, Dresden University of Technology, D-01062 Dresden, Germany, [2]Institute for Biochemistry, Genetic and Microbiology, University of Regensburg, 93053 Regensburg, Germany, [3]BioArCh, Department of Biology, University of York, Heslington, York, YO10 5YW, UK, [4]Technology Facility, Department of Biology, University of York, Heslington, York YO10 5YW, UK, [5]Centre of Excellence in Mass Spectrometry, University of York, Heslington, York YO10 5DD, UK, [6]Department of Chemistry, University of York, Heslington, York YO10 5DD, UK, [7]Bruker Daltonik GmbH, Fahrenheitstrasse 4, 28359 Bremen, Germany, [8]Institute of Bioanalytical Chemistry, Center for Biotechnology and Biomedicine (BBZ), Universität Leipzig, D-014103 Leipzig, Germany, [9]Department für Geo- und Umweltwissenschaften & GeoBio-Center^{LMU}, Ludwig-Maximilians-Universität, D-80333 München, Germany, [10]Department of Geobiology, University of Göttingen, D-37077 Göttingen, Germany, [11]Max Planck Institute of Chemical Physics of Solids, D-01187, Dresden, Germany, [12]Leibniz Institute of Polymer Research Dresden and Max Bergmann Center of Biomaterials, D-01005 Dresden, Germany, [13]Centre d'Oceanologie de Marseille, Station marine d'Endoume, Aix-Marseille Universite–CNRS UMR 6540-DIMAR, 13007 Marseille, France, [14]St. Petersburg State University, 199034 St. Petersburg, Russia, [15]Centre 'Bioengineering' Russian Academy of Sciences, 117312 Moscow, Russia, [16]Institute of Chemistry and Applied Ecology, Far Eastern National University, 690650 Vladivostok, Russia, [17]Max-Bergmann Center of Biomaterials and Institute of Materials Science, Dresden University of Technology, D- 01069 Dresden, Germany, [18]Institute of Solid State Physics, Dresden University of Technology, D- 01069 Dresden, Germany, [19]European XFEL GmbH, 22761 Hamburg, Germany, [20]Alfred Wegener Institute for Polar and Marine Research, D-27570 Bremerhaven, Germany.
*e-mail: hermann.ehrlich@tu-dresden.de; mc80@york.ac.uk

Chapter 6

Conclusion and Discussion

6.1 Advantages of UniColl

6.1.1 Detecting Novel Peptide Hits

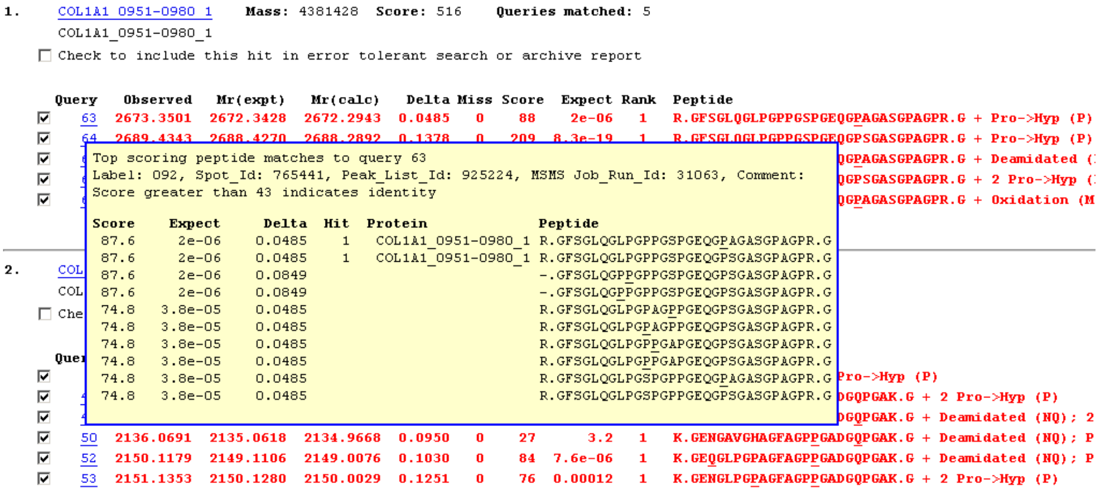
UniColl contains large amount of hypothetical sequences for type I collagen, providing plenty choices for MS/MS database searching. This avoids the difficulty in identifying spectra because of lack of sequences in database. Especially for ancient species and species that are not known in collagen sequences so far, this advantage shows as the matching of novel peptides which have not been discovered in analogous sequences before.

UniColl is more likely to detect novel peptides in unknown sample as a number of sequences not being covered in any conventional protein databases, but are provided with large number of possible matches fit to them in UniColl. The less we know from the sample, the more novel peptides could be discovered. In the case of dodo, since it is homologically close to chicken, whose full type I collagen sequences are contained in most protein databases, the number (four) of novel peptide hits for it is much less than the number (nine) of giant tortoise, which is not close to any known species so far.

6.1.2 Providing All Alternatives

The large number of sequences provides alternative matches for each MS/MS spectrum, including variations on amino acid residues and PTMs. Although this makes peptides matches assigned to incorrect species, all potential information included in MS/MS spectra could be explored.

As shown in Figure 6.1.1, the top 10 alternative matches for each peptide will show in a textbox on the Mascot searching report when licking on the relevant peptide. Information such as matching score, amino acids substitution and PTMs variation for each alternative could be found in the list.



```
1.    COL1A1 0951-0980 1    Mass: 4381428  Score: 516    Queries matched: 5
      COL1A1_0951-0980_1
    □ Check to include this hit in error tolerant search or archive report

     Query   Observed   Mr(expt)   Mr(calc)   Delta Miss Score  Expect Rank  Peptide
   ☑   63   2673.3501  2672.3428  2672.2943  0.0485   0    88     2e-06    1   R.GFSGLQGLPGPPGSPGEQGPAGASGPAGPR.G + Pro->Hyp (P)
   ☑   64   2689.4343  2688.4270  2688.2892  0.1378   0   209   8.3e-19    1   R.GFSGLQGLPGPPGSPGEQGPSGASGPAGPR.G + Pro->Hyp (P)
   ☑    Top scoring peptide matches to query 63                                     QGPAGASGPAGPR.G + Deamidated (
   ☑    Label: 092, Spot_Id: 765441, Peak_List_Id: 925224, MSMS Job_Run_Id: 31063, Comment:   QGPSGASGPAGPR.G + 2 Pro->Hyp (
   ☑    Score greater than 43 indicates identity                                    QGPAGASGPAGPR.G + Oxidation (M

        Score     Expect     Delta   Hit   Protein              Peptide
         87.6      2e-06    0.0485    1   COL1A1_0951-0980_1 R.GFSGLQGLPGPPGSPGEQGPAGASGPAGPR.G
2.  COL  87.6      2e-06    0.0485    1   COL1A1_0951-0980_1 R.GFSGLQGLPGPPGSPGEQGPSGASGPAGPR.G
    COL  87.6      2e-06    0.0849                          -.GFSGLQGPPGPPGSPGEQGPSGASGPAGPR.G
  □ Che  87.6      2e-06    0.0849                          -.GFSGLQGPPGPPGSPGEQGPSGASGPAGPR.G
         74.8    3.8e-05    0.0485                          R.GFSGLQGLPGPAGPPGEQGPSGASGPAGPR.G
         74.8    3.8e-05    0.0485                          R.GFSGLQGLPGPAGPPGEQGPSGASGPAGPR.G
         74.8    3.8e-05    0.0485                          R.GFSGLQGLPGPPGAPGEQGPSGASGPAGPR.G
     Quer74.8    3.8e-05    0.0485                          R.GFSGLQGLPGPPGAPGEQGPSGASGPAGPR.G
   ☑     74.8    3.8e-05    0.0485                          R.GFSGLQGLPGSPGPPGEQGPAGASGPAGPR.G Pro->Hyp (P)
   ☑     74.8    3.8e-05    0.0485                          R.GFSGLQGLPGSPGPPGEQGPSGASGPAGPR.G DGQPGAK.G + 2 Pro->Hyp (P)
   ☑                                                                            DGQPGAK.G + Deamidated (NQ); 2
   ☑   50   2136.0691  2135.0618  2134.9668  0.0950   0    27       3.2    1   K.GENGAVGHAGFAGPPGADGQPGAK.G + Deamidated (NQ); P
   ☑   52   2150.1179  2149.1106  2149.0076  0.1030   0    84   7.6e-06    1   K.GEQGLPGPAGFAGPPGADGQPGAK.G + Deamidated (NQ); P
   ☑   53   2151.1353  2150.1280  2150.0029  0.1251   0    76   0.00012    1   K.GENGLPGPAGFAGPPGADGQPGAK.G + 2 Pro->Hyp (P)
```

Figure 6.1.1 - List of alternative matches for peptide 2673.3501

6.1.3 Assistance of *De Novo* Sequencing

De Novo sequencing is a method of identifying peptides from their MS/MS spectra without the assistance of database searching (Pevtsov *et al.* 2006). It is an effective way to sequence unknown peptides that are not included in databases, except it is time-consuming when calculating all possibilities of sequences generated from mass values of spectra.

UniColl can supply a powerful assistant for *De Novo* sequencing, by generating all probable sequences (based upon known sites of substitution) matched to relevant spectra. This provides alternative selections for *De Novo* sequencing, making it easier to generate the sequence.

Comparing with Asara's claim of *T-rex* sequences in 2007, using only two reference collagens for assistance of *De Novo*, UniColl contains up to 6×10^{211} sequences in the frame of reference. Considering the total number of known vertebrate species which is approximate 4.5 x 10^{4}, UniColl is likely to cover a large percentage of all the vertebrate type I collagen sequences. This makes it a powerful assistance of *De Novo* sequencing.

6.1.4 Peptide-Specific Collagen Database

UniColl is a peptide-specific type I collagen database, different from species-specific databases, it is an expert in identifying type I collagen from the peptide view. Without the restriction of species assignment, it is effective to mine information for every peptide in unknown samples such as ancient species, and mixed samples such as meat-and-bone meal. Also UniColl is collagen-specific database, filtering all non-collagen sequences in searching, could get more accurate matches for collagen samples than protein databases such as UniProt.

6.2 Revelations of UniColl

6.2.1 Types of Misidentification

In this investigation, peptide matches were judged by their mascot ion score. Some spectra could generate a list of alternative sequences with equal ion scores. If the sequence corresponding to relevant species is included in the list, it will be selected as the answer. However if the proper answer gains a lower score, it is called misidentification.

Several types of misidentifications were found very commonly in type I collagen samples of human and cow investigated in the test.

The first type is the inversion of two amino acids which caused by the absence of individual peaks indicating to the correct order of them in spectrum, such as ‘GAPG’ identified as ‘GPAG’. This kind of misidentification can be marked as ‘GA/PG’. The most common inversions of this type are ‘P/A’ and ‘S/P’, suggesting that the chemical bonds of proline-alanine and serine-proline are harder to cleave in mass spectrometer (Martin 1998). Another inversion does not happen to two amino acids next to each other, but very close to each other with several internal amino acids. Take the most popular P/A inversion as an example, ‘PGEA’ identified as ’AGEP’ will be marked as ((PA)GE(AP)), or ‘P/~/A’ in which ‘~’ could stand for one to three amino acids in middle of them. Examples for this kind of inversion misidentifications found in fresh samples test are listed below.

‘P/A’, ‘S/P’, ‘P/~/A’, ‘T/~/A’, ‘V/~/A’

The second type, also a more common type, is misidentifying certain amino acid to another one with identical mass value. Some amino acids and their modified residues are identical in mass, such as lsoleucine (I) and leucin (L), oxidation Methionine (M^) and Phenylalanine (F), hydroxyproline (P*) and L or I, deamidated glutanmine (Q”) and glutamine acid (E). They would be marker as (LI) and (P*L) etc. Examples for this type in the investigation are listed below.

‘(LI)’, ‘(P*I)’, ‘(P*L)’, ‘(M^F)’, ‘(Q”E)’

This type also happens in groups when alternative series of amino acids have the similar mass value to the mass of this fragment. For example, ‘GLTGAP’ could be identified as ‘GSTGPP’ because these two fragments are very close in mass. This kind of group would be marked as ‘G((LS)TG(AP))P’, or ‘(LA~SP)’ where ‘~’ stands for zero to three amino acids in middle of L and A or S and P. Examples for group-confusion of type two misidentifications are listed below.

‘(P*A~SP)’, ‘(P*S~SI)’, ‘(LA~SP)’, ‘(IA~SP)’, ‘(NP*~LN)’, ‘(VQ~LN)’, ‘(P*D~VE)’, ‘(PT~AQ”)’, ‘(P*E~NQ)’, ‘(P*K*~TR)’

The third type, similar to the second one, is also misidentifying an amino acid (group) to a wrong sequence. The difference is these amino acids (groups) are not identical in mass. This is caused by critical shortage of information in MS/MS spectra. All examples for this type found in the test are listed below.

‘(AP)’, ‘(SP)’, ‘(PV)’, ‘(IQ)’, ‘(QL)’, ‘(IN)’, ‘(VN)’, ‘(DE)’, ‘(EN)’, ‘(MA)’, ‘(NP)’, ‘(PP~AQ)’, ‘(AN~SA)’, ‘(PA~AV)’, ‘(SQ~AV)’, ‘(SV~NA)’, ‘(DG~GT)’, ‘(AA~SS)’, ‘(AA~PS)’, ‘(VN~MP)’

The frequent patterns of misidentifications listed above give information of amino acids that are likely to be misidentified in MS/MS database searching in UniColl.

6.2.2 New Approach of Reporting Sequences

It is necessary to describe the reliability of sequences reported from mass spectrometry data analysis. Except using ion scores that describing the overall matching degree for spectra to sequences, it is important to illustrate which part of the sequence is reliable, and which part contains alternative options that could possibly be misidentified.

UniColl provides much more options than other protein databases for MS database searching, with large number of probable matches to be analyzed. In comparison of all possible matches, the common parts of all alternatives with high scores should be confirmed as the certain part in sequences, while the different parts suggest uncertainty in sequences.

Based on the types of misidentifications listed above, a special way to report sequences with clarification of all uncertainties in it was developed. This reporting method of peptide sequence could distinguish confident and unconfident parts in sequences, with showing the alternatives for uncertainty.

For example, one peptide is matched to sequence ‘GAPGEPGATGPPGK’, however with alternatives on the 2nd and 3rd residues where could also be 'PA', while the 6th residue has alternative of being 'Q', also the 8th 'A' and the 11th 'P' could be exchanged. In this case, the sequence would be reported with all possible alternatives displayed as ‘GA/PGE(QP)G((AP)TG(PA))PGK’.

When looking into the peptide in Figure 6.1.1, the score for the top ten matches is 87.6 for the first four matches, and 74.8 for the following six ones. As they all have high matching scores, it is difficult to decide which one is the right answer. In order to report the top four matches with equally highest scores as an answer with expressing the uncertain parts and corresponding alternatives, they can be reported as

'GFSGLQG(LP)PGPPGSPGEQGP(AS)GASGPAGPR'.

6.3 Problems of UniColl

6.3.1 Large Database Size

UniColl includes approximately 4.4×10^{6} hypothetical peptides, and the equivalent of 6×10^{211} theoretical type I collagen sequences in the database, maximizing the possibility to cover the actual sequence for any collagen sample that need identification.

However, the large scale of database inevitably leads to low efficiency of matching, reflected in slower searching speed, higher false positive rate and lower accuracy compared to conventional protein databases. As investigated for fourteen samples in Section 5.5.2, the matching accuracy of UniColl was 50% lower than the conventional protein databases.

The reason causing this problem is that the large amount of hypothetical sequences would necessarily contain many very uncommon combinations that could rarely exist in nature.

There are several ways to solve this problem. Firstly, database searching result from UniColl should be judged not only by the ion scores but also the probability values (explained in Section 3.2.5 and Section 4.7). Matches with high score but very low probability value should be carefully decided by further examining into the spectrum.

Secondly, database size control (demonstrated in Section3.2.7 and Section 4.6) could be applied depending on the demand. If one sample gets too many matches with low probability values, which could prevent the right answers to be displayed, the database size of UniColl would be decreased by excluding part of the lowest probable sequences in the database. This will reduce the interference from uncommon options, but also reduce the probability of indentify uncommon samples such as ancient collagen. Therefore this method should be applied selectively.

Thirdly, sequences with combinations that are chemically or biologically impossible should be cleared from UniColl to avoid improper matches in database searching.

6.3.2 Misidentification

Misidentification is peptides from known sample identified as sequences other than the true sequence of the relevant species. As UniColl provides large number of collagen sequences, compared to other databases which contain far less sequences, it is more likely to generate misidentifications in database searching.

However the existing of misidentification is not a defect of UniColl, but the limitation of matching algorithms. If the misidentified sequence is matched to experimental spectrum with higher ion score than the true sequence, there are three possibilities (i) poor spectra that lack of enough information to identify the spectrum; (ii) the spectrum is actually from the ‘misidentified’ peptide which comes from other proteins; (iii) the spectrum is actually from the true sequence but matched to ‘misidentified’ sequence. In the first two cases, the matching result should be removed from the protein sequence mapping of sample, while in the last case we need to estimate the validity of the scoring system, or to develop additional method to help with selecting proper matches if needed.

As discussed above, misidentification could be caused by poor spectra. The absence of certain ion peaks makes it difficult to identify amino acids component on corresponding positions. Peptide matches generated from very poor spectra are not reliable due to loss of information. However most spectra in this investigation only lack information in fractions, this would results in misidentification only for corresponding amino acids but not the whole protein sequence.

6.3.3 Limitations of Peptide-specific Database

UniColl is a peptide-specific database generated on the base of tryptic peptide units in collagen, since the position of tryptic cleavage sites, arginine (R) and lysine (K) are remarkably consistent through most species especially for mammals. Peptides in UniColl are not formatted in the form of protein chains, but grouped by their tryptic cleaved peptide units. This provides a convenient set of references for sequences alignment in UniColl, while makes it unable to identify missed cleavages and amino acid absence in peptides.

If there is a missed cleavage site in a peptide fragment, that means there are two tryptic units in that peptide, which should be separated into two if being cleaved. This kind of peptide can be indentified in species-specific database such as UniProt, as the full protein sequence is in database, which certainly include two or more peptide connected together. However in the case of UniColl, each sequence contains peptides from the same tryptic unit but in different combinations. Therefore no matches will be hit for missed cleavages.

The same thing happens on amino acids absence in peptides. Although type I collagen sequences are highly identical, there are variations on each residue through different species, and amino acid absence is one of the situation. In conventional species-specific databases, the absences are recorded for the corresponding species. However in UniColl, every amino acid residue in collagen is filled with all variations observed on it except absence. Due the irregular sequence coverage rate of the data source for generating UniColl, some of which only covered several peptide fragments but leave lot of blanks in sequences, the amino acid absence in sequences can not be distinguished from these uncovered blanks. Therefore, including absence as one of the variations in each residue will cause large amount of improper combinations generated for UniColl, that is why the absence has been ignored in the hypothetic sequences.

6.4 Potential Future Research

Considering UniColl has series of problems discussed above, it could be improved in the future in the following aspects.

First, further research could be applied on the chemical and biological algorithms in amino acids combination especially for collagen. In this study, patterns of combinations that are unavailable in nature could be excluded from UniColl to improve the database searching accuracy and efficiency.

A second way to improve the accuracy is to reasonably include common missed cleavages and amino acid absences. In order to realize this, a data analysis should be applied on type I collagen data source, to located the frequent missed cleavages and absence present in most of the sequences. These new variation could then be added to the hypothetical sequences in UniColl to increase the coverage of the database.

The UniColl-like database could be developed for proteins other than type I collagen. Although not all types of protein are suitable in this form, which requires their sequences have limited variations and consistent cleavage sites, some proper options such as type III collagen and osteocalcin can be considered.

Some further application would be expanded on the basis of data similarity investigation and collagen patterns analysis. Incorporating with pattern recognition algorithms, patterns for collagen in MS, in sequence, or in hydroxylation could be explored to indentify collagens especially type I collagen.

Last but not least, the UniColl database should be regularly updated with the latest sequenced type I collagen chains supplemented.

List of Appendix

Appendix 1, the alignment of COL1A1 chains, in txt file.

Appendix 2, the alignment of COL1A2 chains, in txt file.

Appendix 3, the fragmentation analysis for COL1A1 and COL1A2, in excel file.

Appendix 4, R program coding, in doc files.

Appendix 5, theoretic sequences for COL1A1 in UniColl, in txt file.

Appendix 6, theoretic sequences for COL1A2 in UniColl, in txt file.

Appendix 7, statistical analysis on 70 tryptic units for COL1A1, in excel file.

Appendix 8, statistical analysis on 65 tryptic units for COL1A2, in excel file.

Appendix 9, statistical analysis on theoretical sequences for COL1A1, in excel file.

Appendix 10, statistical analysis on theoretical sequences for COL1A2, in excel file.

Note: Appendixes involved in this work are available in the CD submitted together with the thesis.

List of References

Mayr E, Linsley EG, and Usinger RL. 1953. *Methods and principles of systematic zoology.* New York: McGraw-Hill.

Sarich VM, and Wilson AC. 1967. Rates of albumin evolution in primates. *National Academy of Sciences* 58(1):142-8.

Fitch WM. and Margoliash E. 1967. Construction of phylogenetic trees. *Science 155*: 279-284.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Systematic Biology 19* (2): 99-113.

Mount DM. 2004. *Bioinformatics: Sequence and Genome Analysis (2nd ed.).* Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.

Stuart AJ. 1975. The vertebrate fauna of the type Cromerian. *Boreas 4*: 63–76.

Stuart AJ. 1982. *Pleistocene vertebrates in the British Isles.* London: Longman.

Abelson PH. 1954. Organic constituents of fossils. *Carnegie Institute of Washington Yearbook 53*: 97–101.

De Jong EW, Westbroek P, Westbroek JF. and Bruining JW. 1974. Preservation of antigenic properties of macromolecules over 70 Myr. *Nature 252*: 63–64

Westbrock P, van der Meide PH. and van der Wey-Kloppers JS *et al.* 1979. Fossil macromolecules from cephalopod shells: Characterization, immunological response and diagenesis. *Paleobiology 5*: 151–167.

Paabo S, Higuchi RG, and Wilson AC. 1989. Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. *Journal of Biological Chemistry 264* (17): 9709-9712.

Schweitzer MH. 2004. Molecular paleontology: some current advances and problems. *Annales de paléontologie 90* (2):81-102.

Eglington G, and Logan GA. 1991. Molecular preservation. Philosophical Transactions of the Royal Society of London Series B. *Biological Sciences 333*:315-328.

Bachmann A, Kiefhaber T, Boudko S, Engel J, and Bächinger HP. 2005. Collagen triple-helix formation in all-trans chains proceeds by a nucleation/growth mechanism with a purely entropic barrier. *Proc Natl Acad Sci U S A 102*(39):13897-902.

Millard A. 2001. Deterioration of bone. In: Pollard AM, and Brothwell D, (Eds.) *Handbook of archaeological sciences*. Chichester: John Wiley and Sons.

Karas M, and Hillenkamp F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry 60*(20):2299-2301.

Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, and Yoshida T. 1988. Protein and polymer analyses up to m/z 100,000 by laser ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry 2* (8):151-153.

Nielsen-Marsh CM, Richards MP, Hauschka PV, Thomas-Oates JE, Trinkaus E, Pettitt PB, Karavanic I, Poinar H, and Collins MJ. 2005. Osteocalcin protein sequences of Neanderthals and modern primates. *Proceedings of the National Academy of Sciences 102* (12):4409-4413.

Collins MJ, Nielsen-Marsh CM, Hiller J, Smith CI, Roberts JP, Prigodich RV, Weiss TJ, Csapo J, Millard AR, and Turner-Walker G. 2002. The survival of organic matter in bone: A review. *Archaeometry 44*:383-394.

Collins MJ, Riley MS, Child AM, and Turner-Walker G. 1995. A basic mathematical simulation of the chemical degradation of ancient collagen*. Journal of Archaeological Science 22*:175- 183.

Asara JM, and Schweitzer MH. 2008. Response to comment on "Protein sequences from mastodon and tyrannosaurus rex revealed by mass spectrometry". *Science 319* (5859):33d.

Asara JM, Schweitzer MH, Freimark LM, Phillips M, and Cantley LC. 2007. Protein sequences from mastodon and tyrannosaurus tex revealed by mass spectrometry. *Science 316* (5822):280-285.

Bowman S. 1990. *Radiocarbon Dating*, ISBN: 9780520070370

Nobuhiro Zaima, Takahiro Hayasaka, Naoko Goto-Inoue, and Mitsutoshi Setou. 2010. Matrix-assisted laser desorption/ionization imaging mass spectrometry. *International Journal of Molecular Sciences 11*(12):5040-55.

Pappin DJ, Hojrup P, and Bleasby AJ. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol. 3*(6): 327–32

Hume, Julian Pender (2006). The History of the Dodo Raphus cucullatus and the Penguin of Mauritius. *Historical Biology 18* (2): 69–93

Pevtsov S, Fedulova I, Mirzaei H, Buck C,and Zhang X. 2006. Performance evaluation of existing de novo sequencing algorithms. *J Proteome Res 5*(11):3018-28.

Martin RB. 1998. Free energies and equilibria of peptide bond hydrolysis and formationl. *Biopolymers 45:* 351–353.

Laursen RA. 1971. Solid-phase Edman degradation: An automatic peptide sequencer. *Eur J Biochem.11*; 20(1):89-102 ().

Mann M. and Wilm MS. 1994. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem. 66:* 4390-4399 (1994).

Eng, J. K., McCormack, A. L. and Yates, J. R. 1994. An approach to correlate MS/MS data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom. 5:* 976-989.

Perkins DN, Pappin DJ Creasy DM. and Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis 20:* 3551-3567.

Aebersold R. and Goodlett DR. 2001. Mass spectrometry in proteomics. *Chem. Rev 101:* 269-295.

Mann M, Hendrickson RC, and Pandey A. 2001. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem. 70*: 437-473

Aebersold R. and Mann M. 2003. Review article: mass spectrometry-based proteomics. *Nature 422*: 198-207

Richard S. Johnson, Michael T. Davisa, J. Alex Taylorb and Scott D Pattersona. 2005. Informatics for protein identification by mass spectrometry. *Methods. Mass Spectrometry in Proteomics. 35:* 223-236