

# Protein structure comparison: The quest for a meaningful measure of similarity

*Toby Lane<sup>1</sup>, Peter Whigham<sup>2</sup>*

<sup>1</sup>Spatial Information Research Centre  
University of Otago, Dunedin, New Zealand  
Phone: +64 3 479-8301 Fax: +64 3 479-8311  
Email: tlane@infoscience.otago.ac.nz

<sup>2</sup>Spatial Information Research Centre  
University of Otago, Dunedin, New Zealand  
Phone: +64 3 479-7391 Fax: +64 3 479-8311  
Email: pwhigham@infoscience.otago.ac.nz

Presented at SIRC 2002 – The 14<sup>th</sup> Annual Colloquium of the Spatial Information Research Centre  
University of Otago, Dunedin, New Zealand  
December 3-5<sup>th</sup> 2002

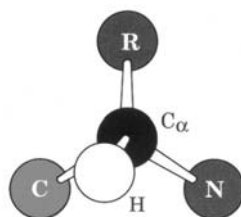
## ABSTRACT

To a biochemist there is perhaps nothing so elusive as the three dimensional structure of a favourite protein. As such the development of a fast and reliable method to computationally determine protein structure is something of a holy grail, and presents one of the major challenges to science. In the assessment of the reliability of a given prediction technique, it is necessary to quantitatively compare the differences between a predicted output, and the desired one. Here we examine the shortcomings of some traditional structure comparison techniques, as well as presenting the challenges brought by new methods.

**Keywords and phrases:** protein structure prediction, RMS, similarity, wavelet

## 1.0 INTRODUCTION

Proteins are typically described as the workhorses of the cell, and are essential for all biological processes. They have very defined roles and chemical makeup in order to carry out these roles in the most efficient manner possible. As such they have a well-defined three-dimensional structure. Proteins are composed of amino acids, the precise order of which is defined by the gene that encodes the protein. Four chemical groups are bound to the central alpha carbon. These are the amino group, the carbonyl group, hydrogen, and the R group, which defines the properties of the residue (fig 1).



*Figure 1: General structure of amino acid in three dimensions, showing the relative spacing of the side groups.  
Modified from (Donald Voet 1995)*

Proteins are built on the ribosome in a process called translation, where tRNA molecules toting single amino acids bind briefly with the mRNA from the nucleus. The tRNA molecules give up their amino acid load, which is then bound to the previous amino acid in a condensation reaction. The bond between the amino acids is called

a peptide bond. This bond has partial double bond characteristics, making rotation about this bond unlikely. Consequently the atoms surrounding the peptide bond ( $\omega$ ) are in an almost planar arrangement (fig 2). The result is that the major rotation in a chain of residues occurs within the amino acid about the  $\phi$  and  $\psi$  angles.

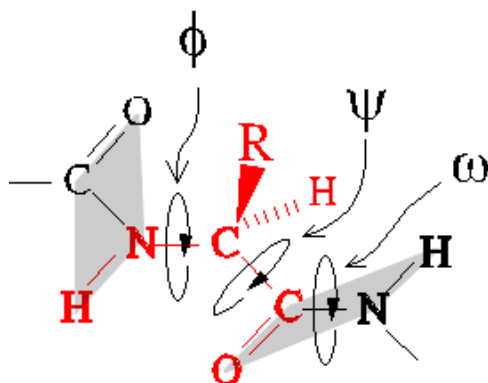


Figure 2: Possible rotations within the peptide backbone,  $\phi$  and  $\psi$  are the major contributors. Grey areas highlight planar  $\omega$  bonds.

Rotation about the  $\phi$  and  $\psi$  angles is restricted by the side chains (R-group), and their steric interactions. The side chains themselves define the characteristics of the amino acid, and the sum of these characteristics determines those of the final protein structure. Amino acids may be basic, acidic, polar, non-polar and charged or non-charged. They may fluctuate between these states depending on their environment.

This stepwise addition of amino acids continues until eventually the full sequence of the protein is completed. The next step in the life of a protein is folding. Exactly how the primary sequence finds its final three-dimensional shape, out of the many possible conformations is not completely understood. Since Levinthal presented his interesting paradox (Levinthal 1968), we have known that the process is not a random sampling of conformations. Further study through the years has revealed that many factors are involved. These factors are the physical forces that, in conjunction with covalent interactions, determine a protein structure, and its folding pathway. Hydrophobic interactions are generally considered to be the most important of these forces. It is in fact the absence of hydrogen bonding between water and the internal non-polar residues that is important for the stability of globular proteins, and provides a driving force for folding (Nölting 1999). The final structure of the protein determines its function, and it is the amino acid sequence, or primary structure that dictates structure. The native state of the protein, or its conformation when carrying out its physiological role is usually the conformation of most interest. It should be noted that this structure is not a rigid one; in fact the dynamic nature of protein structure is integral to protein functionality. Representing this dynamic structure is a non-trivial problem, and will become increasingly more important for future biochemical and drug development studies. The native state is a thermodynamically stable structure in physiological conditions, but is not necessarily the most energetically minimized structure possible. Hydrogen bonding also has an important role within a protein molecule, particularly in stabilizing secondary structural elements. Other forces such as van der Waals interactions and dipole-dipole forces also contribute to the stability of protein tertiary structure. Protein structure is generally broken down into 4 levels. Primary structure is the sequence of amino acids that make up the protein. Secondary structure is the local structure, which contains defined elements such as alpha helices and beta sheets. The discovery of many chaperone proteins has further complicated the protein-folding story. Chaperones assist protein folding by preventing, and reversing unfavourable conformations.

In the quest to discover the method by which proteins fold, many mechanisms have been proposed, perhaps the most well known of these is the molten globule theory. The original molten globule theory is attractively simple; its mechanism involves the collapse of the polypeptide chain, a process driven by hydrophobic interactions. The original model assumed a lack of long-range interactions found in the native conformation (Privalov 1996). This model has been further refined, and in its original form largely dismissed as a kind of “black box” for protein folding. The revised versions, such as the framework model and nucleation condensation mechanism (figure 3) hint that the formation of long range interactions in the transition states is necessary for the correct formation of the native conformation (Nölting 2000). However the extent to which the interactions form, and precisely at what point in the pathway they form is still largely contentious.

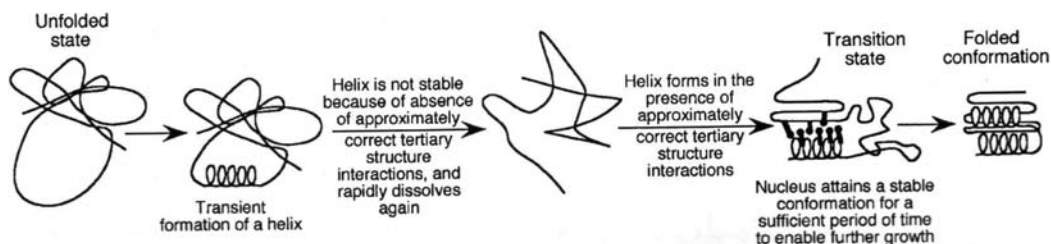


Figure 3: Folding pathway, showing importance of long-range interactions (nucleation-condensation mechanism) (Nölting 2000).

## 1.1 Structural analysis

### 1.1.1 Biophysical methods

Traditionally the most successful tool for finding protein structures is X-ray crystallography. This technique involves growing crystals of a protein; the crystal is then exposed to a beam of x-rays. The x-rays are diffracted, and detected by an x-ray film (figure 4).

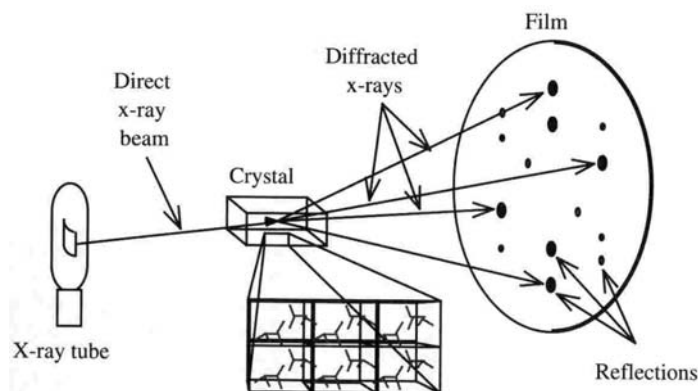


Figure 4: System for crystallographic data collection (Rhodes 1993)

Atoms in the crystal diffract the x-rays, and the resulting pattern is specific to the molecule. This pattern can be analysed to produce an electron density map, which corresponds to the structure of the protein molecule. It is the process of growing proteins that is the limiting factor in x-ray crystallography. Protein crystals, unlike inorganic ones, are held together by weak forces, such as hydrogen bonds. As a result, protein crystals are very fragile, and need to be treated accordingly. Also the actual formation of crystals is something of a dark art, needing a fine balance of many factors including pH, temperature, concentration of protein and precipitant. These combine to make the process of protein crystallisation slow and often disappointing.

Nuclear Magnetic Resonance, or NMR, is a newer tool than x-ray crystallography, and can often produce higher resolution images. Because the proteins must be in solution, it has the added benefit of avoiding the rigour of protein crystallisation. NMR is limited, however, to small proteins. Due to the nature of the technology, the maximum protein size depends largely on the strength of the magnetic coil in the NMR machine. As the technology improves, the upper limit will only increase.

### 1.1.2 Computational methods

Because of the difficulties in finding protein tertiary structures through physical methods, many researchers have attempted to develop computer systems for predicting protein folding. Systems have attempted brute force techniques with some recent success, notably the folding@home program run by Stanford University (Shirts and Pande 2000), (Snow 2002). The folding@home program follows in the footsteps of the popular SETI@home program. By taking advantage of spare processor time donated by willing participants, this brute force approach has been able to solve some small peptide structures to within an acceptable error. Many artificial intelligence approaches have also been tried, such as knowledge based systems, hidden Markov Models, and particularly neural networks (Wohlfahrt, Hangoc et al. 2002), (Karplus, Karchin et al. 2001), (Petersen, Lundegaard et al. 2000). These approaches have had varying degrees of success.

Artificial neural networks are biologically inspired systems, which are made up of processing elements (neurons), which are arranged in layers. These neurons are interconnected, and each of these connections has a weighting, which may or may not be trainable. They are designed to have some of the characteristics of the human brain: learning and adaptation, generalisation, parallelism and robustness (Kasabov 1996). With the development of the Multilayer Perceptron (MLP) and algorithms sophisticated enough to train them, neural networks have become more popular, and many new architectures have been developed.

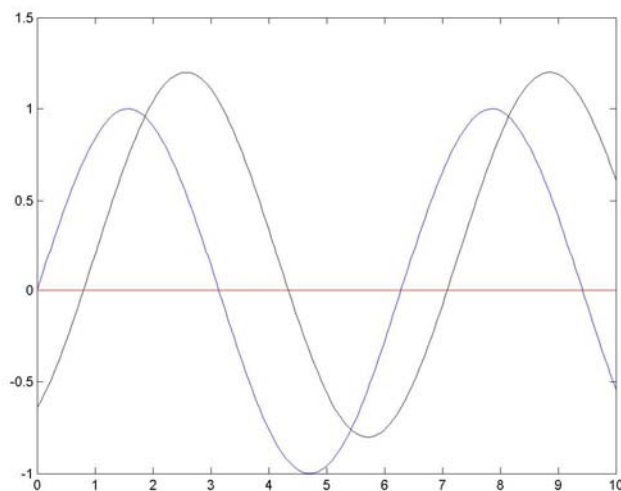
Any system for discovering protein structure, be it computational or biophysical, has some measure of error that is used to direct the development of the model. For computational systems, particularly neural networks, that error is critical, and will determine the accuracy of the system. How that error is calculated, so that a meaningful and appropriate response can occur, is essential for improvements in such a prediction system.

## 2.0 MEASURING ERROR

Any supervised learning system requires the ability to assess the quality of its outputs, so that outputs can be compared. This error must be quantifiable, so that outputs can be ranked in order of fitness. This is essential for improvement. It is also important that this assessment be a valid measure, which represents the important features of the system being modelled. It must also recognize not just when an exact match has been made, but also where similar features are found in different areas of space, that is it must be able to cope with the two signals being out of phase. It should also identify common trends over the whole of the two structures being compared, as well as the localised zones. This measure must recognise not only the areas where the system is failing, but also the areas where the system is achieving, and must then weight these failures and achievement in a valuable and meaningful way.

### 2.1 RMS Error

Traditionally root mean square (RMS) error has been used for measuring protein similarity (Mirny and Shakhnovich 1998). This has been carried over from crystallography uses into systems for structure prediction. However it does not fit the desired characteristics for an error measure technique as discussed above. RMS is a point-based comparison technique, and as such does not take into account trends, or motifs in a data structure. This is highlighted in figure 5 below, which shows three curves. Two of these are sine curves, and one a linear curve. However RMS calculations show a higher level of similarity between the linear curve and a sine curve, than between the two out of phase sine based curves.



*Figure 5: Two sine curves have greater difference than either sine curve to the straight line according to RMS. RMS clearly has some faults.*

Other commonly used point-based techniques, such as absolute percentage error and relative absolute error, are limited by the same problem. Some efforts have been made to use other methods for calculating error, which attempt to take into account themes and trends within a signal.

## 2.2 Fourier analysis

Fourier analysis has been used with great success in many scientific fields. The basic assumption of the Fourier process is that any periodic function can be represented as the sum of a collection of sines and cosines. This treatment allows seemingly complex patterns to be broken down and their underlying formulae exposed for inspection. By exposing these patterns, it is then possible to see the properties and characteristics that define a particular model.

Fourier transform has several uses including image searching through databases as well as image compression (Gibson and Gaydecki 1995). Unfortunately it has some drawbacks. Fourier transforms were designed for discovering patterns in periodic systems, and as such are not well suited to non-periodic systems. Because it breaks a signal down into a collection of periodic signals, information about the location of an event in the original signal is dispersed throughout all frequencies of the transform. This has the effect of hiding sudden changes within the original signal. Often it is these sudden changes that give the most information about the original signal, for instance a border in an image, or in a protein a change from alpha helix to irregular coil. Windowed Fourier transform offers some solution to this problem. By passing a window of analysis over a signal sudden changes can be identified, however the compromise is reduced understanding of the overall themes and trends in the signal. Wavelets however can be used to overcome these problems.

## 2.3 Wavelets

The history of wavelets is not as long, nor as clear as that of Fourier transform. The recent increase in popularity of wavelets has largely been made possible by the availability of computing power. Wavelet analysis is an extension of Fourier analysis, and as such shares many characteristics. The basic principle of comparing the wavelet to the original signal to produce coefficients is the same, however rather than using a collection of periodic functions to represent the system, wavelet analysis stretches and compresses the non-periodic wavelet function and compares them to the original signal at each frequency. The result is a collection of coefficients that describe the relationship between the wavelet and the signal at each level of detail. The coefficients can then be compared, and a measure of similarity obtained. Unlike Fourier methods, wavelets allow all levels of detail to be represented. A particular improvement over Fourier transform is the ability to observe sudden changes in a signal, while simultaneously capturing larger trends (figure 6). This process has been used successfully for image recognition and search algorithms (Wang J. Z 1997). This record in image recognition, as well as an ability to discern between high-level themes and low-level detail, champions wavelets as a method for error measurement in structure comparison.

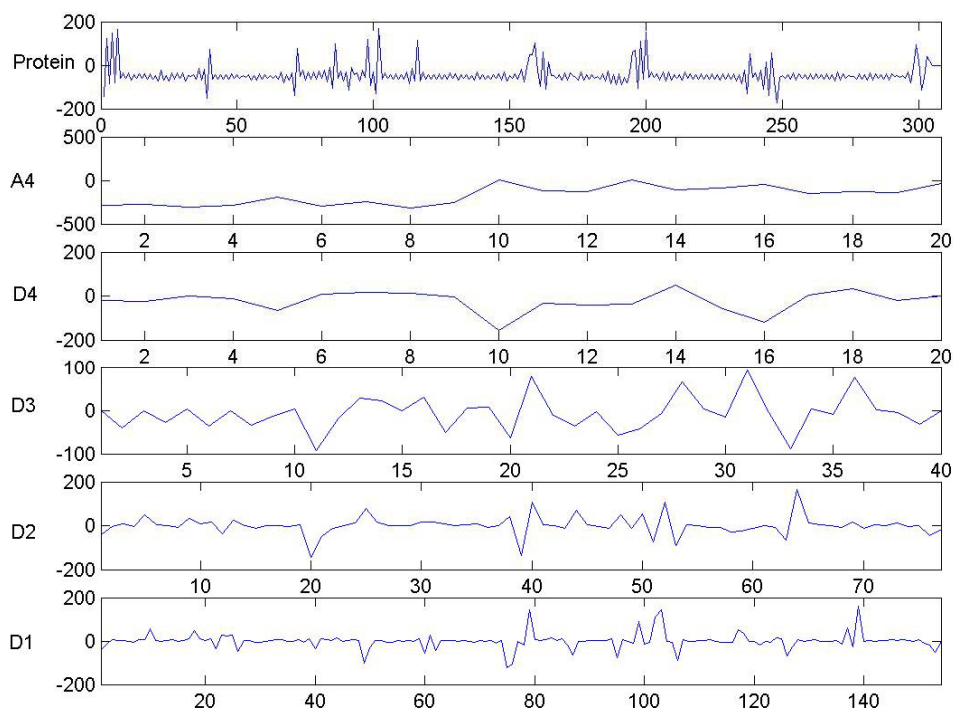


Figure 6: Discrete four layer wavelet decomposition of protein phi-psi signal from sperm whale myoglobin PDB ID 109m (Berman, Westbrook et al. 2000). D4 though D1 show increasing levels of detail over four scales.

### 3.0 METHODS AND RESULTS

Data were collected from the Protein Data Base (Berman, Westbrook et al. 2000). The PDB files were converted to phi-psi angles using the MOLEMAN program from Uppsala Software Factory (Kleywegt 1997). They were imported into MatLab (version 6.1), and plotted as a signal (see figure 6 phi-psi signal). Following the work of Wang *et al.* (Wang J. Z 1997) we used a discrete wavelet transform over four layers on each of the signals with the Daubechies 1 wavelet (Daubechies I. 1992). The difference between the layers was calculated, and each layer difference multiplied by a weighting factor. These products were then added together to produce a final error measure. The weighting factor for each layer was calculated as follows: each layer in turn was set to a matrix of zeros, and then the signal was reconstructed. The difference between the original signal and the reconstructed one was calculated for each layer removal. These weights were then normalised based on the highest weight and used as weighting factors.

In order to evaluate this method several test cases were prepared based on protein 109m. The first test case swapped residues 131 – 140 with 141 – 150. These were chosen, as they are part of a sequence of alpha helical structure. The second test case replaced the ten residues from 21 – 30 with random phi and psi values. Also a set of alterations was randomly generated to view the relationship between change and wavelet similarity. This was done by randomly choosing a residue within the 154 amino acid protein chain, and changing it to a random value within the allowed Ramachandran domains (Ramachandran G. N. 1963). This was repeated so that 154 changes were made. Both the wavelet similarity and RMS were calculated between these new signals and the original. The results are plotted in figure 7.

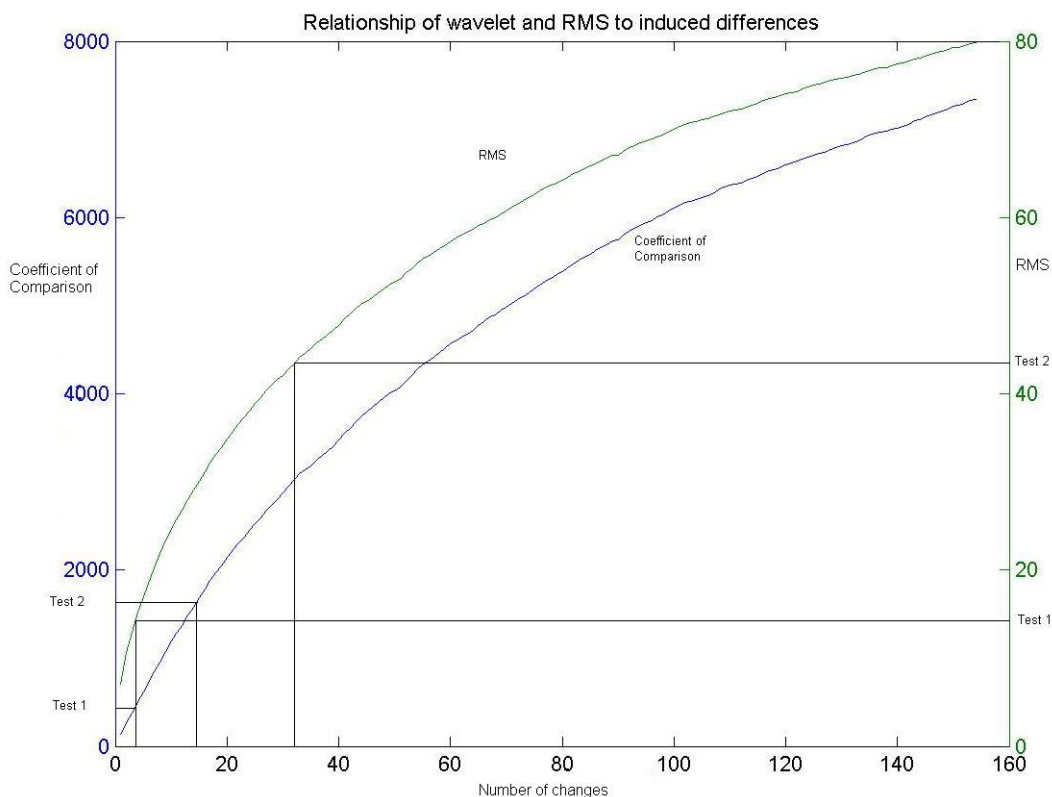


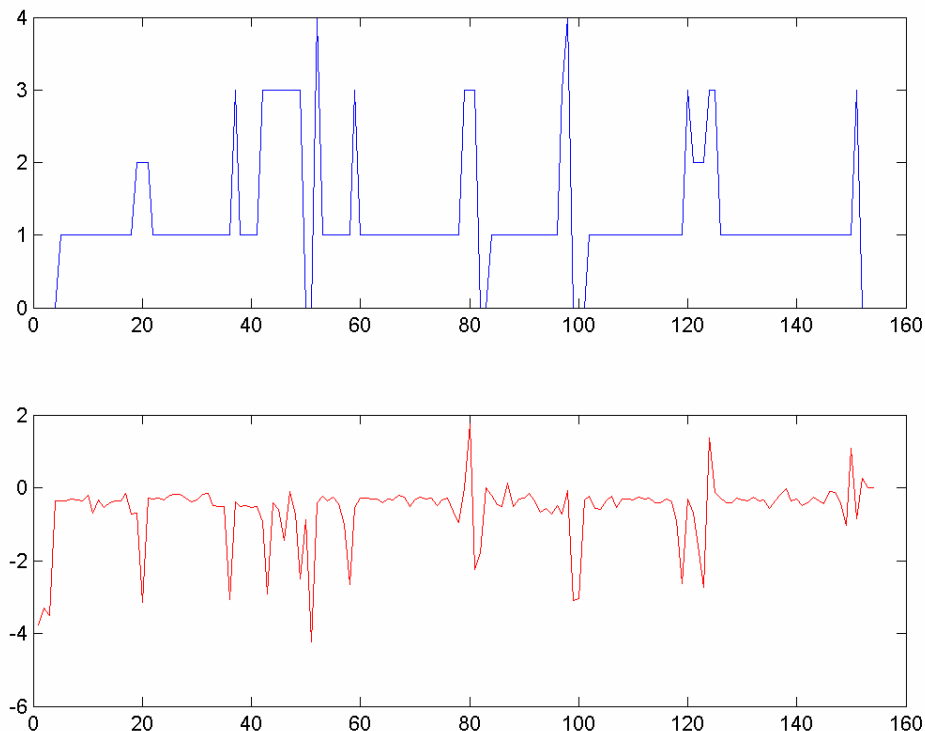
Figure 7: This curve shows both the RMS and wavelet similarity results for the stepwise modification. Also plotted are the results from test cases 1 and 2. This result shows the discrepancy between wavelet analysis and RMS analysis.

Figure 7 shows a comparison between wavelets and RMS. Assuming wavelets were a more appropriate measure of similarity, Test 1 and Test 2 should have shown comparatively less error for wavelets than RMS. However, it is clear that both measures are performing in a similar fashion. Hence it does not appear that wavelet descriptions are capturing the higher-level structural changes that were expected.

However, it was noticed that wavelets were effective at identifying changes in structure, that is the boundaries between secondary structural elements, as shown in Figure 8. Since secondary structure is a fundamental description of protein structure this is a positive result that showed some wavelet characteristics that did capture higher-level properties, and suggests that wavelet analysis can be used as a method for feature extraction. To do this positions and numbers of peaks, as well as their sign need to be taken into account. This provides many challenges for future development. A possible method for capturing this information is to compare the positions of elements within the signal, and assign a positive score for every match. Elements could be considered to be any change in the signal of greater than one standard deviation between points. Matching could be done with a simple binary output for a match, and calculate a percentage score for the number of matches. This score is a less than ideal, as it does not bring about a meaningful measure of comparison. It fails to identify the occurrence of similar structural elements if they are not aligned, so may have difficulty recognising themes if they occur out of phase. It would be possible to also count the numbers of each significant feature, but the relative importance of the two values would vary between cases. The importance will also vary depending on the exact task at hand, and how similarity is defined. For instance more significance may be placed on secondary structure similarity than overall tertiary structure, or vice versa. Indeed relative importance may even change during the course of a training exercise, as different areas of a model are refined. This change could potentially be exploited in a genetic algorithm, so that the measure of fitness changes over time. It may be effective to generate a population of neural networks that produce similar numbers of similar features, with their exact locations in space ignored. This seed population could then be exposed to the selection pressure of alignment and ordering of features and themes.

It may also be worthwhile to include expert knowledge into the measure of similarity, so that parameters can be adjusted in real time.

Another possible line of research would be to consider using a classifier-approach that modelled structure by using many individuals in an evolving population to solve separate portions of the problem (Holland 1978). A final solution would then be constructed by combining these individuals. The argument for this approach is that the problem is too complex to be represented or solved by one model, and may be more tractable with a divide-and-conquer approach.



*Figure 8: The secondary structure elements of protein 109m (top) and the layer 1 wavelet decomposition (bottom). Note the correspondence between changes in secondary structure and peaks in the layer 1 wavelet signal.*

## 4.0 CONCLUSION

It seems clear that similarity measures are far from perfect. We have highlighted the shortcomings of the traditional point based measures such as RMS, and suggested possible improvements, which harness wavelet technology. In the process the subjective nature of similarity has become apparent, and in the end the result may be that we need to apply a separate measure for each situation.

## REFERENCES

- Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." *Nucleic Acids Res* **28**(1): 235-42.
- Daubechies I. (1992). *Ten lectures on wavelets*, SIAM.
- Donald Voet, J. G. V. (1995). *Biochemistry*, John Wiley & Sons, Inc.

- Gibson, D. and P. A. Gaydecki (1995). "Definition and application of a fourier domain texture measure: applications to histological image segmentation." Comput Biol Med **25**(6): 551-7.
- Holland, J. R., J. (1978). Cognitive systems based on adaptive algorithms. Pattern-directed Inference Systems. F. H.-R. D.A. Waterman. New York, Academic Press.
- Karplus, K., R. Karchin, et al. (2001). "What is the value added by human intervention in protein structure prediction?" Proteins **45**(Suppl 5): 86-91.
- Kasabov, N. K. (1996). Foundations of neural networks, fuzzy systems, and knowledge engineering. Cambridge, Mass., MIT Press.
- Kleywegt, G. J. (1997). "Validation of protein models from Calpha coordinates alone." J Mol Biol **273**(2): 371-6.
- Levinthal, C. (1968). "Are there pathways for protein folding?" J. Chim. Phys. **85**: 44-45.
- Mirny, L. A. and E. I. Shakhnovich (1998). "Protein structure prediction by threading. Why it works and why it does not." J Mol Biol **283**(2): 507-26.
- Nölting, B. (1999). Protein folding kinetics : biophysical methods. Berlin ; New York, Springer.
- Nölting, B. A. k. (2000). "Mechanism of Protein Folding." Proteins: Structure, Function, and genetics **41**: 288-298.
- Petersen, T. N., C. Lundegaard, et al. (2000). "Prediction of protein secondary structure at 80% accuracy." Proteins **41**(1): 17-20.
- Privalov, P. L. (1996). "Intermediate states in protein folding." J Mol Biol **258**(5): 707-25.
- Ramachandran G. N., R. C., Sasisekharan V. (1963). "Stereochemistry of Polypeptide Chain Configurations." J. Mol. Biol **7**: 95-99.
- Rhodes, G. (1993). Crystallography made crystal clear : a guide for users of macromolecular models. San Diego, Academic Press.
- Shirts, M. and V. S. Pande (2000). "Screen savers of the world unite." Science **290**(5498): 1903-1904.
- Snow, C. N., H. Pande, V. Gruebele, M. (2002). "Absolute comparison of simulated and experimental protein-folding dynamics." Nature advance online publication.
- Wang J. Z, W. G., Firchein O, Wei S. X (1997). Wavelet-based Image Indexing Techniques with Partial Sketch retrieval Capability. Proc. 4th Forum on research and Technology Advances in Digital Libraries.
- Wohlfahrt, G., V. Hangoc, et al. (2002). "Positioning of anchor groups in protein loop prediction: The importance of solvent accessibility and secondary structure elements." Proteins **47**(3): 370-8.