

Prediction of Genome and Protein Structures of *Homo Sapiens* Neanderthalensis Using NCBI Tools

Rakesh Medarametla^{1*}, Lavanya Rao² and Kishore Bejugam³

¹School of Life sciences, University of Wolverhampton, UK

²Department of Biochemistry, Dr. L. B College, Visakhapatnam, AP, India

³Mallareddy College of pharmacy, Hyderabad, India

Abstract

The national Centre for biotechnological information (NCBI) is part of the United States of America. The role in research and education specific to fundamental molecular and genetic processes that control health and disease is done by National Library of Medicine (NLM) NCBI. Introduction to NCBI databases and maps help researchers select tool that may help in finding sequences. A particular helpful feature is science primer, which provides an easy to read introduction to many science topics relevant to NCBI resources like Bio informatics, Molecular modelling, Genome mapping and Molecular genetics. An effort to sequence entire genome adds to the number of protein sequences by a factor of two each year, the gap between sequence and structural information stored in Public databases is growing rapidly. In contrast to sequencing techniques, an experimental method for structure determination takes lot of time and limited in their application. Therefore it's not able to keep pace with the flood of newly characterised gene products. The growth of practical methods for predicting protein structure from sequence is therefore of significant importance in field of biology.

Keywords: Genome; Protein structure; *Homo sapiens*

Introduction

Neanderthals should be classified as *Homo neanderthalensis* or "*Homo sapiens neanderthalensis*". DNA consists of phosphates, sugars and bases arranged in a chain. There are four different bases: adenine, thymine, guanine and cytosine. DNA is wound together in a double stranded molecule with adenine paired with thymine and cytosine paired with guanine across the two sides of the strand. When DNA is replicated, the chemical bonds that bind together the two sides are broken and new nucleotides are joined to the unwound strand, generating strands with the same sequences as the original. Mutations arise when an error in the copying occurs and a different base is substituted. Different combinations of bases code for different amino acids, which are the building blocks of proteins. Sometimes more than one combination of bases leads to the same amino acid. For example, both AAA and AAG code for lysine. Therefore, a DNA change that led to a G at the third position rather than an A wouldn't affect the code for lysine. This is known as a silent change. However, if the third position had changed to a C, making the set AAC that would change the amino acid to asparagine. This change that could affect the protein is called

non synonymous. Most human sequences differ from each other by on average 8.0 substitutions, while the human and chimpanzee sequences differ by about 55.0 substitutions. The Neanderthal and modern human sequences differed by approximately 27.2 substitutions. Mitochondrial DNA from the Paglicci specimens as well as other ancient humans fit within the range of modern humans, but the Neanderthals remain consistently genetically distinct. This shows that early anatomically modern *Homo sapiens* were not very different genetically from current modern humans, but were still different from Neanderthals. Though this evidence does not disprove the idea of Neanderthal and modern human admixture, it shows that moderns and Neanderthals did not have more genetic similarities during the Pleistocene that were subsequently lost [1].

Prediction of Gene Structure

In above Figure 1, gene is present on the chromosome 1. On chromosome 1, gene is present in region 1p36.1. The protein encoding this gene is member of GTP binding protein. CDC42 gene represented in dark red colour transcribed from 3' to 5' direction (Figure 2) [2].

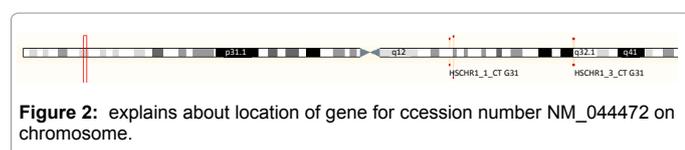
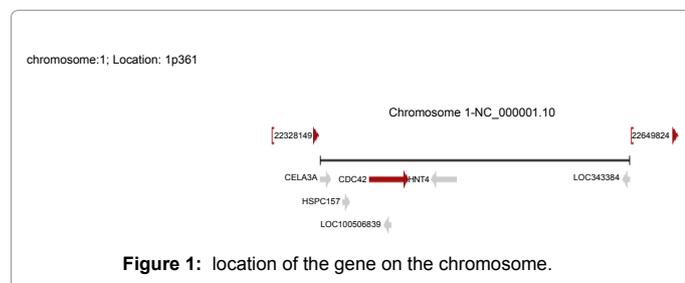
Genomic Regions, Transcripts and Products

Figure 3 explains about genomic regions, transcripts and products [3].

Protein Structure Prediction

Secondary structure of protein

In structural biology and biochemistry, the secondary structure can



*Corresponding author: Rakesh Medarametla, School of Life sciences, University of Wolverhampton, UK, Tel: 9553602253; E-mail: rakesh.medarametla2253@gmail.com

Received April 15, 2014; Accepted April 18, 2014; Published April 20, 2014

Citation: Rakesh M, Lavanya R, Kishore B (2014) Prediction of Genome and Protein Structures of Homo Sapiens Neanderthalensis Using NCBI Tools. J Data Mining Genomics Proteomics R1: 001. doi:10.4172/2153-0602.R1-001

Copyright: © 2014 Rakesh M, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

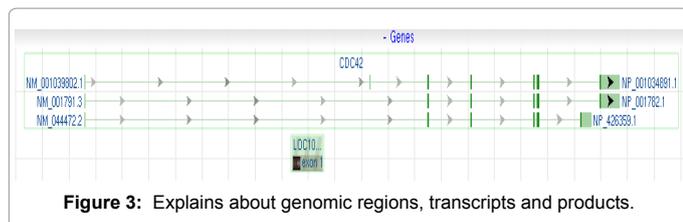


Figure 3: Explains about genomic regions, transcripts and products.

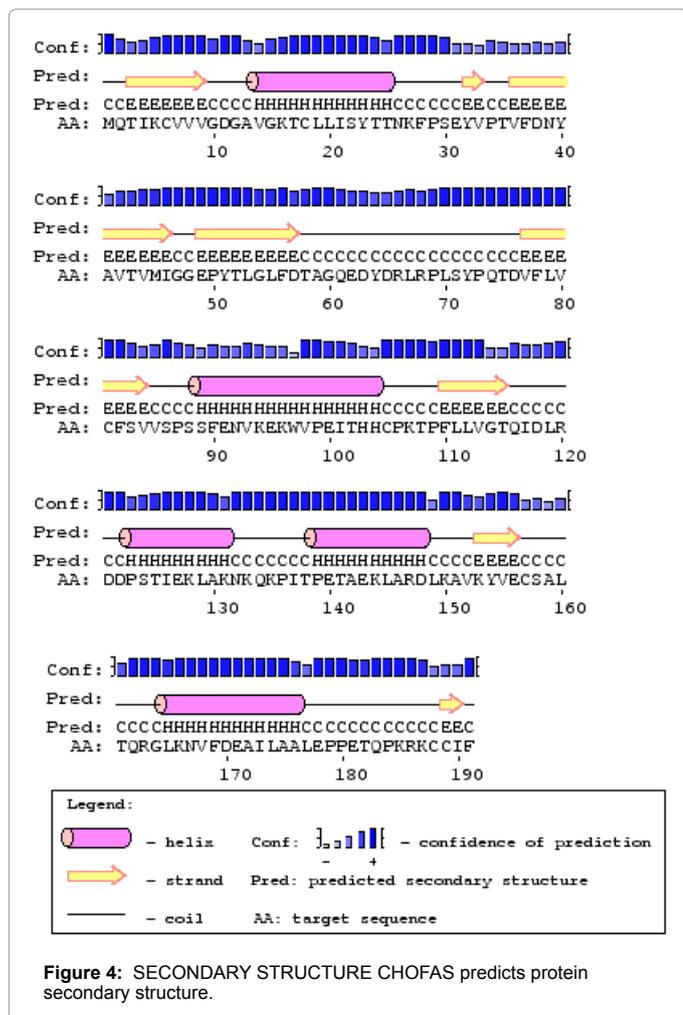


Figure 4: SECONDARY STRUCTURE CHOFAS predicts protein secondary structure.

be defined as the general three dimensional forms of local segments of biopolymers such as nucleic acids (DNA or RNA) and proteins. To know the relation between the amino acid sequence and protein structure an unambiguous and physically meaningful definition of secondary structure is vital [4-8].

The explosive accumulation of protein sequences in large scale sequencing projects are in severe contrast to much slower experimental determination of protein structures. The improved methods of predicting the structure from the gene sequence alone are therefore needed [8-12].

The structure of proteins can be predicted in four types.

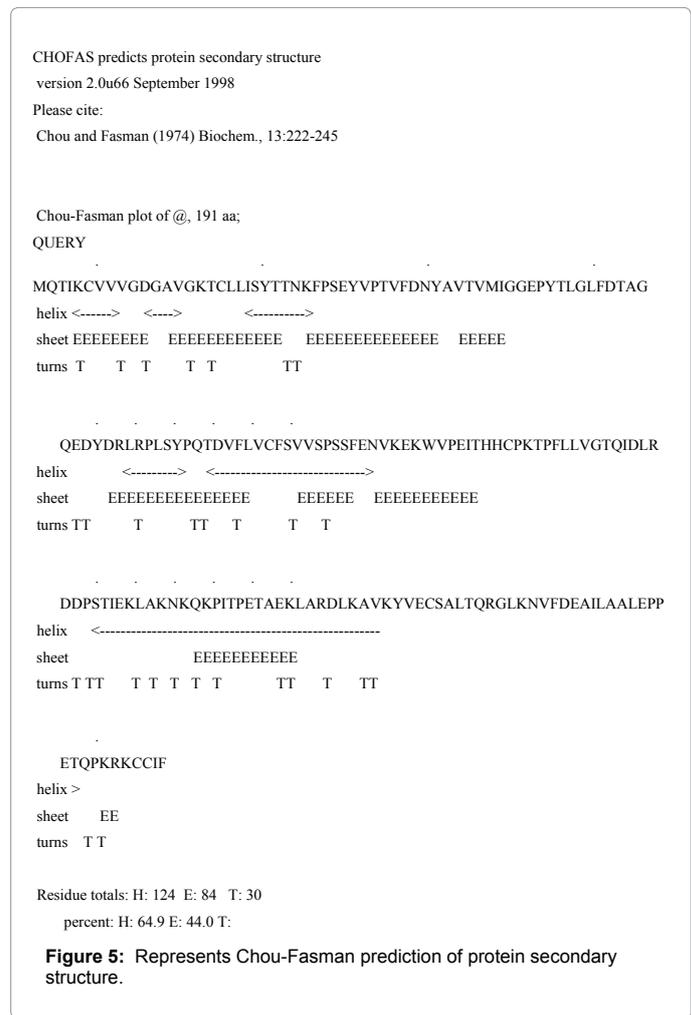
1. Proteins are made up of amino acids arranged in poly peptide chains, and the order of the amino acids in this chain is known as primary structure.

2. The regular way in which the poly peptide chains are arranged in space to form a protein is called secondary structure.
3. The arrangement of three dimensional structure of protein in space is known as tertiary structure.
4. The arrangement of combination of two or more poly peptides chains is known as Quaternary structure.

Secondary structure of protein can be predicted by using the website <http://bioinf.cs.ucl.ac.uk> in these protein FASTA sequence was uploaded the results appears as follows.

Figure 4, determines the graphical representation of protein produced from psipred server.

The prediction of secondary structure (pred) is observed in three different colours (Legend) for three different types of secondary structure. Pink colour cylindrical legend represents helix, yellow colour arrow represents strand, black colour line represents random coil of secondary structure. The confidence of prediction (conf) is indicated in blue colour vertical bars. Each blue bar is prediction for each amino acid. Greater the height of bar, darker the blue colour, higher the confidence of prediction level. Smaller the bars, paler the blue colour then lower the confidence of prediction level. From Figure 4, confidence of prediction in secondary structure is mostly high and Helix, coil and β stands are shown to be in equal composition.



CHOFAS predicts protein secondary structure
version 2.0u66 September 1998
Please cite:
Chou and Fasman (1974) Biochem., 13:222-245

Chou-Fasman plot of @, 191 aa;
QUERY
MQTIKCVVVG DGAVGK TLLISY TTNKFPSEYVPTVFDNYAVTVMIGGEPYTLGLFDTAG
helix <-----> <-----> <----->
sheet EEEEEEEEE EEEEEEEEEEE EEEEEEEEEEEEE EEEEE
turns T T T T T TT

QEDYDRLRPLSYPTDVFVCFVSVVSPSSFENVKEKVVPEITHHCPKTPFLVGTQIDLR
helix <-----> <----->
sheet EEEEEEEEEEEEEEE EEEEE EEEEEEEEEEE
turns TT T TT T T T

DDPSTIEKLAKNKQKPIPTAEKLRDLKAVKYVECSALTQRGLKNVFDEAILAALAPP
helix <----->
sheet EEEEEEEEEEE
turns TTT T T T T TT T TT

ETQPKRCCIF
helix >
sheet EE
turns TT

Residue totals: H: 124 E: 84 T: 30
percent: H: 64.9 E: 44.0 T:

Figure 5: Represents Chou-Fasman prediction of protein secondary structure.

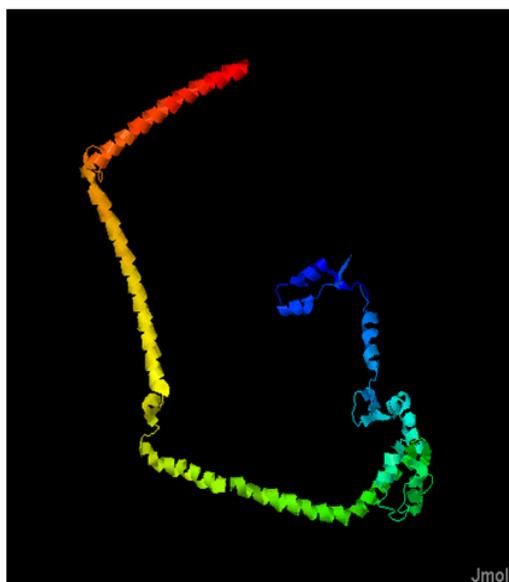


Figure 6: Tertiary structure of accession number NM_044472.

Secondary Structure of Protein Prediction

Secondary structure is a set of techniques in bioinformatics that aims to predict secondary structure of proteins and nucleic acid sequences based on the basic information of primary structure.

For analysis of secondary structure Chou-Fasman plot, (3 corresponding) psipred server, (colour coded short corresponding states) Spom method (colour coded states with graphical view correspond) were used.

In Figure 5, secondary structure of protein predicted with basic model Chou-Fasman plot. The above sequence is protein FASTA format of given accession number NM_044472 composed with total residues of 191 amino acids. Below FASTA sequence, residue of secondary structure are represented in helix (<---->), Sheets (E) and turns (T). Of the total 191 residues, helix with 124 residue (64.9%), Sheets with 84 residue (44.0%) and turns with 30 residues shown respectively.

Chou-fasman prediction is a basic secondary structure when compare to psipred prediction. Chou-fasman predicts helix, sheet, and turns for each residue with over all percentage but does not show confidence of prediction in graphical representation.

Tertiary Structure

The accurate secondary-structure prediction is an important aspect for predicting the tertiary structure. The tertiary structure is a unique three dimensional folding of a protein. The tertiary structure is considered to be largely determined by the sequence of amino acids of which it is composed and by the proteins primary sequence. The protein tertiary structure is held together by interactions between the side chains- the "R" groups. The hydrogen bonding, hydrophobic interaction, ionic bonding and disulphide bonds are the forces that give rise to the tertiary structure of a protein. By using <http://www.sbg.bio.ic.ac.uk> protein structure prediction server allows us to submit a protein sequence. It then performs a prediction of our choice and receives the results of the prediction via e-mail [12-16].

To understand the tertiary structure, we should also know β -sheet location and also the regions with no regular α -helix and about the secondary structure fit in [17,18].

The study of figure can provide a very good idea of the underlying structural framework of tertiary contacts between alpha-helices, in both globular and TM environments. The Figure 6 shows included distances, angles, and dihedral angles that together describe entirely the underlying geometry of each contact. From the fig it is likely to identify differences in the geometry requirements of tertiary contacts between alpha-helices in common and Trans membrane alpha-helices. Finally, how the geometry of tertiary contacts changes with the amino acid can be studied [19-21].

Raster Display of Molecules (RasMol) is a molecular graphics program used for the visualization of nucleic acids, small molecules and proteins. This program is more important for teaching, display and generation of publication quality images [22-24].

References

1. <http://en.wikipedia.org/wiki/Neanderthal>.
2. <http://www.ncbi.nlm.nih.gov/Taxonomy/browser/wwwtax.cgi>
3. <http://www.ncbi.nlm.nih.gov/blast.cgi>
4. Alberts B (2003) Essential Cell Biology. New York, NY: Garland Science.
5. Nakata K (1995) Prediction of zinc finger DNA binding protein. *Bioinformatics* 11: 125-131.
6. Pauling L, Campbell DH, Pressman D (1943) The nature of the forces between antigen and antibody and of the precipitation reaction. *Physiol Reviews* 23: 203-219.
7. Barnwal RP, Chary KVR, (2008) An efficient method for secondary structure determination in polypeptides by NMR. *Current Science* 94: 1302-1306.
8. Meiler J, Baker D (2003) Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci U S A* 100: 12105-12110.
9. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433: 128-132.
10. Furnham N, Blundell TL, DePristo MA, Terwilliger TC (2006) Correspondence: is one solution good enough. *Nature Struct and Mol Biol* 13: 184-185.
11. Rother D, Sapiro G, Pande V (2008) Statistical characterization of protein ensembles. *IEEE/ACM Trans Comput Biol Bioinf* 5: 42-55.
12. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J (2000) Prediction of protein secondary structure at 80% accuracy. *Proteins* 41: 17-20.
13. Chou PY, Fasman GD (1974) Prediction of protein conformation. *Biochem* 13: 251-276.
14. Lim VI (1974) Structural principles of the globular organization of protein chains: A stereo-chemical theory of globular protein secondary structure. *J Mol Biol* 88: 857-872.
15. Rose GD (1978) Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* 272: 586-590.
16. Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 120: 97-120.
17. Maxfield FR, Scheraga HA (1976) Status of empirical methods for the prediction of protein backbone topography. *Biochem* 15: 5138-5153.
18. Robson B, Suzuki E (1976) Conformational properties of amino acid residues in globular proteins. *J Mol Biol* 107: 327-356.
19. Taylor WR, Thornton JM (1983) Prediction of super-secondary structure in proteins. *Nature* 301: 540-542.
20. Rooman MJ, Kocher JP, Wodak SJ (1991) Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *J Mol Biol* 221: 961-979.

21. Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol* 214: 171-182.
22. Zhang X, Mesirov JP, Waltz DL (1992) Hybrid system for protein secondary structure Prediction. *J Mol Biol* 225: 1049-1063.
23. Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1992) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J Mol Biol* 212: 151-166.
24. Goldman N, Thorne JL, Jones DT (1996) Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J Mol Biol* 263: 196-208