Protein structure prediction and structural annotation of proteomes

Book or Report Section

Accepted Version

Roche, D. B., Buenavista, M. T. and McGuffin, L. J. (2018) Protein structure prediction and structural annotation of proteomes. In: Encyclopedia of Biophysics. Springer. doi: https://doi.org/10.1007/978-3-642-35943-9_418-1 Available at http://centaur.reading.ac.uk/79209/

Gordon C. K. Roberts
Encyclopedia of Biophysics
10.1007/978-3-642-16712-6_418

Protein Structure Prediction and Structural Annotation of Proteomes

Daniel Barry Roche[1], Maria Teresa Buenavista[2, 3, 4] and Liam James McGuffin[2]
(1) Institut de Biologie Computationnelle, LIRMM, CNRS-UMR 5506, Université de Montpellier, 34095, Montpellier, France

(2)School of Biological Sciences, University of Reading, Whiteknights, Harborne Building, Reading, RG6 6AS, UK

(3)Biocomputing, MRC Harwell, Harwell Oxford, Oxfordshire, OX11 0RD, UK

(4)Diamond Light Source Ltd., Beamline B23, Chilton, Didcot, OX11 ODE, UK

Daniel Barry Roche (Corresponding author)
Email: daniel.roche@lirmm.fr

Maria Teresa Buenavista
Email: m.buenavista@har.mrc.ac.uk

Liam James McGuffin
Email: l.j.mcguffin@reading.ac.uk

Abstract

Proteins are essential molecules for the functioning of all living organisms. Thus, studying the structure of proteins enables us to better determine their function. Experimental methods for structural determination, namely X-ray crystallography and NMR, are timely and are costly. Hence, a more efficient method for structure determination are computational methods.

Structural annotation of proteomes, where the structure of all proteins in a proteome are determined is more routinely undertaken, as whole genome sequencing, using NGS technology, of large numbers of organisms have been undertaken. The gap between protein sequence and experimental structure is widening at an exponential rate. Thus, computational methods for protein structure prediction are essential to help close this gap.

Synonyms

Fold recognition; Protein structure; Protein structure modeling; Protein structure prediction; Sequence alignments; Structural genomics; Template-based modeling; Template-free modeling

Definition

Protein structure prediction methods aim to predict the structures of proteins from their amino acid sequences, utilizing various computational algorithms. Structural genome annotation is the process of

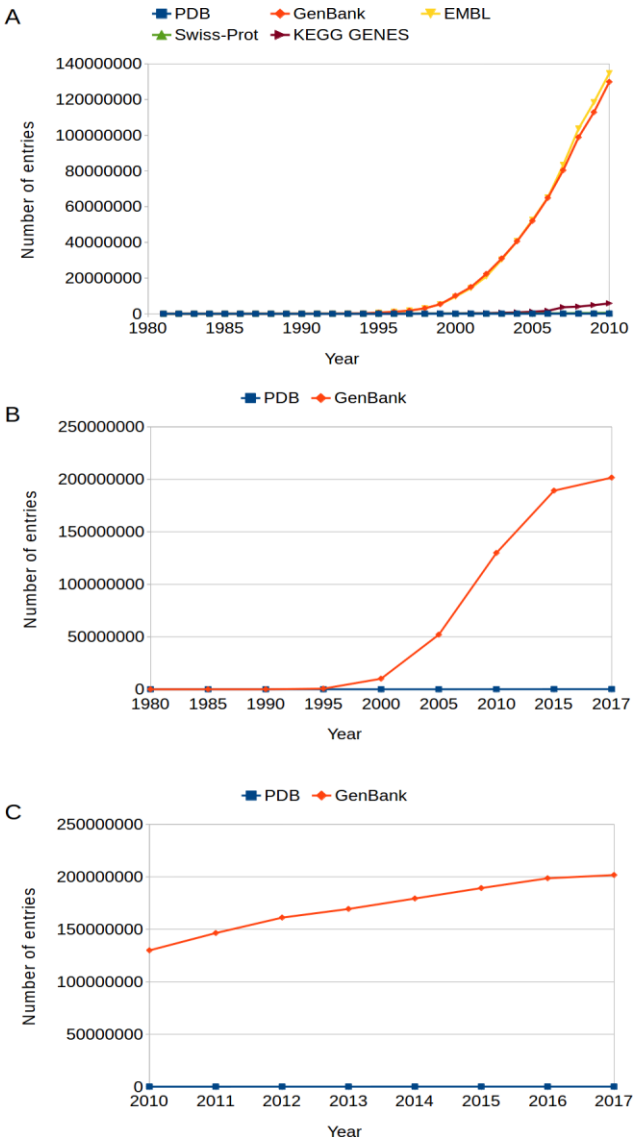
attaching biological information to every protein encoded within a genome via the production of three-dimensional protein models.

Introduction

Proteins are essential molecules involved in both structural and functional roles of all living cells. Numerous diseases, notably Alzheimer's, Parkinson's, heart disease, and cancers, involve mutations in specific proteins, which affect their function. Thus, determining protein structure is essential to understanding functionality and is potentially helpful in developing treatments for the diseases or disorders. The ultimate goal is to determine the function of a protein from sequence computationally, but often sequence information alone is insufficient. During the course of evolution, protein structure has been more conserved than amino acid sequence, therefore the analysis of protein structures often leads to a greater understanding of protein function than can be obtained from just studying their sequences (McGuffin *2008a*).

Experimental methods which include X-ray crystallography and nuclear magnetic resonance (NMR) are commonly used for protein structural determination, but they have several limitations. The cloning, expression, and purification of a protein and in the case of the X-ray crystallography, the subsequent production of diffraction quality crystals for protein structure determination, are time consuming and costly. Conversely, computational methods for protein structure prediction are easily automated, fast, and cheap. Predicted structures allow the inference of function, lead to a better understanding of protein evolution, and guide experimental work in drug discovery, biopharmaceuticals, industrial enzymes, ligand–protein interactions, and cancer biology, to name a few applications.

In this post-genomic era, the gap between protein sequence and structure is widening. At the time of writing, there are <133,000 protein structures in the Protein Data Bank (PDB) and ~200 million sequences in the GenBank database (Fig. 1). The rate at which 3D structures are being solved is evidently unable to compete with the speed of genome sequencing. However, the use of bioinformatics tools may be used to help close the gap between sequence and structure, help in proteome annotation, and speed up the elucidation of protein structures by the production of high quality homology models for molecular replacement.



If you need to edit the image, please use the original: 978-3-642-16712-6_17_Part_Fig1-418_HTML.gif

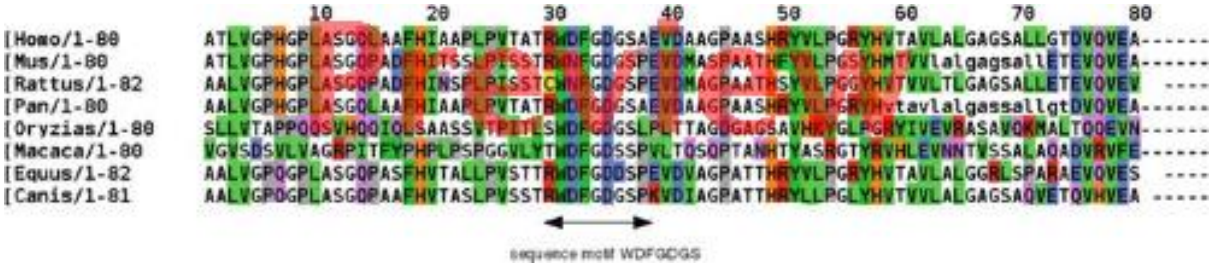
Protein Structure Prediction and Structural Annotation of Proteomes, Fig. 1

The number of sequences deposited in sequence databases including GenBank, EMBL, Swiss-Prot, and KEGG GENES is dwarfing the number of protein structures in the PDB by a factor of over 2000

(Data taken from the PDB and sequence databases). (**a**) The number of sequences in sequence databases and the number of protein structures in the PDB are plotted against the years of entry, just focusing on the PDB and GenBank. (**b**) As in (**a**) but updated to take into account the growth between 2010 and 2017. (**c**) The number of sequences in sequence databases and the number of protein structures in the PDB is plotted against the years of entry, focusing on the growth between 2010 and 2017, for the PDB and GenBank.

Sequence Alignment in Protein Structure Prediction

Sequence alignment algorithms are subdivided into global and local sequence alignment methods. Global sequence alignment algorithms seek to align two sequences over the whole length of the protein to produce a score, which determines the evolutionarily relatedness of the two proteins. Local sequence alignment algorithms align evolutionarily related segments of proteins, which include binding sites, domains, and sequence repeats that are important to the protein, but may exist in other proteins, which are distantly evolutionarily related and have unrelated functions (Fig. 2) (Altschul et al. *1990*, *1997*; Lipman and Pearson *1985*).



If you need to edit the image, please use the original: 978-3-642-16712-6_17_Part_Fig2-418_HTML.jpg

Protein Structure Prediction and Structural Annotation of Proteomes, Fig. 2
Multiple sequence alignment of the PKD1 domain 1 (PDB 1B4R), showing highly conserved sequence motif WDFGDGS. The eight species were aligned using ClustalW2, and the HHpred color scheme was utilized to illustrate amino acids with similar biochemical properties

The first global sequence alignment algorithm was developed by Needleman and Wunsch in 1970 (Needleman and Wunsch *1970*), which was the first application of dynamic programming for sequence comparison. This was followed in 1981 by Smith and Waterman's (Smith and Waterman *1981*) development of an algorithm for local sequence alignment possessing a high degree of similarity, which included the addition of a weighting for gap penalties and the use of a matrix to identify sequence pairs.

Lipman and Pearson developed FASTA (Lipman and Pearson *1985*) in 1985 with an update to the algorithm in 1988. FASTA is a local sequence alignment method, which is still widely used. FASTA was one of the first sequence alignment algorithms that could be run on a standard desktop PC of the era, through the introduction of the concept of "ktups." Ktups are segments of an amino acid sequence, with ktup = 1 being one amino acid long, ktup = 2 is two amino acids long and so forth. This concept is used to increase the speed of the algorithm, as the number of searches is reduced. The higher the ktup value, the faster the speed. FASTA utilizes ktup = 2 as the default for amino acid sequence alignment and ktup = 6 for DNA sequence alignment (Lipman and Pearson *1985*).

BLAST – Basic Local Alignment Search Tool – (Altschul et al. *1990*) is a rapid sequence alignment algorithm for homology searching of sequence libraries. BLAST, like FASTA, also utilizes the ktups method but refers to ktups as "words," with the default "word" size set to three for amino acid

sequences and 11 for DNA sequences. Gapped BLAST and PSI-BLAST (Position Specific Iterative-BLAST), introduced in 1997 (Altschul et al. *1997*), further improved the sensitivity and speed of the BLAST algorithm. Gapped BLAST generates a single gapped alignment, which increases the speed for pairwise sequence alignment, whereas the original BLAST program often finds several alignments involving a single database sequence, which when considered together were statistically significant. Using the Gapped BLAST alignment algorithm, it then becomes necessary to find only one rather than all of the ungapped alignments significantly increasing the speed of the algorithm. PSI-BLAST is more sensitive for the detection of weak, but biologically relevant sequence similarities. PSI-BLAST uses a sequence-profile alignment method, with a position-specific scoring matrix generated from significant alignments in round i which are then used in round $i + 1$ to generate a matrix for the next round. This iterative searching and profile construction process significantly increases the sensitivity of searching the sequence and structure databases. PSI-BLAST is an integral part of most successful tertiary structure prediction pipelines.

Critical Assessment of Techniques for Protein Structure Prediction (CASP)

The continual development of more advanced protein structure prediction tools is driven by the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition. CASP is a biennial competition, with the aim of advancing the methods of predicting protein structures from sequence, by the provision of objective testing of the methods via blind prediction. CASP is currently divided into a number of categories prediction categories: (1) tertiary structure prediction – template-based and free modeling, (2) disorder prediction, (3) domain prediction, (4) contact prediction, (5) quality assessment, (6) binding site prediction and more recently (7) structural refinement (Moult et al. *2009*) (Kryshtafovych et al., 2016).
There have been major improvements seen in the structure prediction category in each successive CASP. The previous three CASP experiments showed that fully automated structure prediction servers can produce models close in quality to those produced by the very best expert human modelers (Kryshtafovych et al. *2009*). Server performance is extremely important, since they are the only choice for high throughput modeling. The number of servers involved in CASP has increased from 53 in CASP5 to 79 in CASP9, additionally 85 in CASP11, showing an increase in interest for protein structure prediction and increased competition in the field.

Tertiary Structure Prediction

Protein tertiary structure prediction methods are divided into template-based and template-free modeling methods. If a structural template is available within the PDB, template-based modeling methods such as homology modeling and fold recognition can be utilized, but if a structural template is unavailable free modeling will have to be utilized (Table 1).**Protein Structure Prediction and Structural Annotation of Proteomes, Table 1**
Established techniques for the modeling of protein folds fall into three major categories, which is dependent on the level of information that is known about the protein sequence (McGuffin *2008b*)

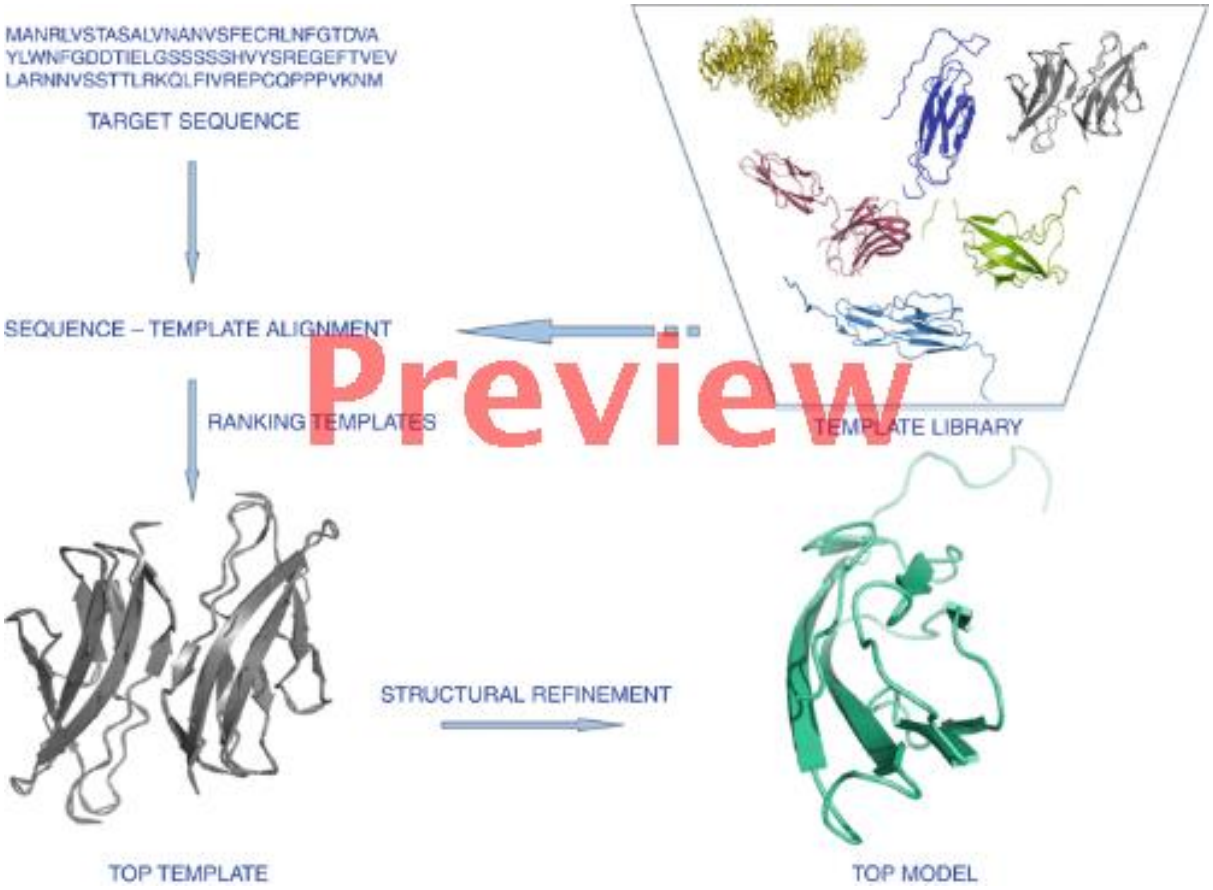
Method category	Requirements	Relative computational	Relative speed	Theoretical sequence

		difficulty		coverage
Homology/comparative modeling	Homologous (>30% sequence ID) to a template structure from the PDB	Easy	Fast	Minimum
Fold recognition/threading	A template fold of known structure from the PDB	Medium	Medium	Medium
Ab intio/new fold/free modeling	The target sequence and/or a fragment library	Hard	Slow	Maximum

Template-Based Modeling

The success of template-based modeling (TBM) methods is based on three key facts: (1) similar sequences fold into similar structures, (2) many unrelated sequences also fold into similar structures, and (3) there are only a relatively small number of unique folds when compared with the number of proteins found in nature; most of the fold space has been structurally annotated and few new folds are being solved (McGuffin *2008b*).

Traditionally, template-based modeling is divided into two subcategories: homology modeling and fold recognition. Homology modeling, also known as comparative modeling, is dependent on finding a sequence alignment between the target and the template structure with a sequence ID > 30%. Fold recognition methods go beyond simple sequence searching, when the sequence identity between the target and template sequence is within the twilight zone (20–35% sequence ID). However, it is becoming increasingly difficult to differentiate between homology modeling and fold recognition algorithms, as most successful methods now utilize profile–profile-based sequence searching algorithms to identify very distant relationships between targets and templates (McGuffin *2008b*). HHsearch is a popular, rapid, and accurate profile–profile-based fold recognition method (Soding *2005*), which utilizes the PSI-BLAST position-specific scoring matrices and PSIPRED (Jones *1999*) secondary structure predictions, to build profile-Hidden Markov Models (HMMs) of the target sequences. The profile-HMMs are then compared to a fold library of profile-HMMs that have been built for proteins with known structure. Once target-template alignments have been constructed, then 3D models of the target structure can be built from the coordinates of the template structures (Fig. 3) (Table 2).



If you need to edit the image, please use the original: 978-3-642-16712-6_17_Part_Fig3-418_HTML.gif

Protein Structure Prediction and Structural Annotation of Proteomes, Fig. 3
Template-based modeling pipeline, such as that used by IntFOLD-TS

Protein Structure Prediction and Structural Annotation of Proteomes, Table 2
Some of the top publicly available template-based modeling servers in CASP11

Server	URL
BioSerf	http://bioinf.cs.ucl.ac.uk/bio_serf/public_job
HHpred	http://toolkit.lmb.uni-muenchen.de/hhpred
IntFOLD	http://www.reading.ac.uk/bioinf/IntFOLD/
I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
LOMETS	http://zhanglab.ccmb.med.umich.edu/LOMETS/
PCONS	http://pcons.net/
Phyre2	http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

pro-sp3-TASSER	http://cssb.biology.gatech.edu/skolnick/webservice/pro-sp3-TASSER/index.html

A user-friendly template-based modeling prediction pipeline – IntFOLD3-TS - is now available which combines profile–profile alignment outputs from several different methods to produce up to 40 alternative fold recognition models, which are subsequently ranked utilizing the ModFOLDclust2 (McGuffin and Roche *2010*) model quality assessment method. The IntFOLD server (Roche et al. *2011*) (McGuffin et al., 2015) also integrates the ModFOLD5 method for model quality assessment, the DISOclust3 method for protein intrinsic disorder prediction, the DomFOLD3 method for domain boundary prediction, and the FunFOLD3 method for the prediction of ligand-binding site residues.

Template-Free Modeling

Free modeling has also been referred to as *ab initio* modeling, modeling from first principles, or *de novo* modeling. Template-free modeling is the prediction of a protein's tertiary structure from sequence without the use of a protein structure as a template. The use of free modeling is necessary when a template cannot be found to predict the structure of the protein. Free modeling usually carries out conformational searches under the assistance of a designed energy function, which generates several structural decoys based on possible conformations that the final model is selected from. Energy functions for free modeling are generally classified into physics-based energy functions and knowledge-based energy functions, which depend on the use of statistics from structurally elucidated proteins (Table 3) (Lee et al. *2009*).**Protein Structure Prediction and Structural Annotation of Proteomes, Table 3**

Some of the top publicly available free modeling web servers in CASP11

Server	URL
TASSER-VMT	http:// http://cssb.biology.gatech.edu/skolnick/webservice/TASSER-VMT/index.html
MULTICOM-NOVEL	http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html
QUARK	http://zhanglab.ccmb.med.umich.edu/QUARK
RAPTORX	http://raptorx.uchicago.edu
Robetta	http://robetta.bakerlab.org/

Physics-Based Methods

Physics-based methods are defined as methods that utilize interactions between atoms based on quantum mechanics and electrostatic interactions. Physics-based methods also utilize a small number of critical parameters, which include electron charge and Planck's constant, with atoms additionally described by their atom type, where only the number of atoms is relevant. The use of quantum mechanics has not yet been used to predict even small structures, due to the computational resources needed for such calculations. Without the use of quantum mechanics, the most practical starting point for free modeling is to utilize a compromised force field, with a large number of selected atom types. Within each atom type the physicochemical properties are calculated from information on crystal

packing or quantum mechanical theory. Examples of all-atom physics-based force fields include AMBER, CHARMM, and OPLS, which also contain terms in relation to bond length, angles, torsion angles, van der Waals, and electrostatic interactions (Lee et al. *2009*).

Knowledge-Based Methods

Knowledge-based methods are generally more successful and utilize empirical energy terms, derived from structurally elucidated proteins deposited in the PDB. These energy terms are further divided into two sub-classifications. The first energy term encompasses genetic and sequence-independent terms, which include hydrogen bonding and the local backbone stiffness of a polypeptide chain. The second energy term encompasses amino acid or protein sequence-dependent information, which include: pairwise residue contact potentials, distance dependent atomic contact potentials and secondary structure propensity. Despite most knowledge-based methods utilizing secondary structure propensities, local structures may be rather difficult to reproduce when modeling. One way in which this problem can be counteracted is the utilization of secondary structure fragments, acquired from sequence or profile alignments, for the initial model construction. This is also advantageous as the entropy of the conformational search is reduced (Lee et al. *2009*). The fragment assembly method for knowledge-based free modeling was utilized in FRAGFOLD (Jones *2001*), ROSETTA and the QUARK servers (Table 3).

Model Quality Prediction

Once a selection of models has been produced for a target sequence, the quality of each model must then be assessed. Being provided with details about the potential errors in 3D models arguably makes them more useful in the context of guiding experimental work. Model quality assessment programs (MQAPs) are used for the prediction of 3D model quality of proteins (McGuffin and Roche *2010*). MQAPs can be classified into two categories: single model–based methods, which are able to assess the quality of individual models, and the clustering-based methods, which compare multiple models against each other. According to recent CASP experiments, the clustering-based MQAP methods, such as ModFOLDclust2 (McGuffin and Roche *2010*), are currently the most accurate methods if multiple alternative models can be obtained. However, the single model methods, such as ModFOLD6 (Maghrabi and McGuffin, 2017), which produce absolute scores for individual models, are potentially more useful if few models are available.

Structural Annotation of Genomes

The structural annotation of genomes is extremely important for the functional determination of the encoded protein sequences. Traditionally, functional annotation of protein sequences has been carried out using simple sequence alignment methods. With the rapid growth in the number of genome projects, the need for accurate annotation has also increased. However, sequence-based structural annotation is inadequate for protein sequences, which have a low pairwise sequence identity (<30%) to proteins with known structures. Thus, structural annotation methods, which attempt to carry out fold recognition on a genomic scale, can help to increase the level of annotation beyond the twilight zone of sequence identity.

Structural Annotation Databases

Several databases have been developed providing structural annotations of genomes, including Gene3D (Yeats et al. *2008*), SUPERFAMILY (de Lima Morais et al. *2011*), and the Genomic Threading Database (McGuffin et al. *2004*). These databases have been constructed via the use of sequence-based searching methods and fold recognition in order to structurally annotate entire proteomes. Gene3D provides an up-to-date comprehensive database for structural and functional annotations of the majority of available protein sequences, including UniProt, RefSeq, and Integr8. Structural annotations of genomes, are generated via a detailed search of the CATH structural database profile-HMM library. Functional annotation is also carried out by the Gene3D database utilizing GO assignments, FunCat, KEGG, active site data, disordered predictions utilizing DISOPRED2, and data from microarray experiments (Yeats et al. *2008*).

Methods for Structural Annotation

Both intensive fold recognition methods such as IntFOLD-TS and rapid methods such as HHsearch (Soding *2005*) can be utilized for structural annotation of entire proteomes. McGuffin et al. structurally annotated the entire human proteome utilizing the mGenTHREADER method (McGuffin and Jones *2003*) in just over 24 h by harnessing 515 CPUs. This study provided the proof of concept that intensive fold recognition can be carried out for rapid proteome annotation via the use of grid technology (McGuffin et al. *2006*). Recently, we undertook the structure annotation of 1,838,675 proteins from over 600 genomes, derived from metagenomic or single sequencing (Roche and Bruls, 2015). This showed that it is now possible to carry out large-scale structural annotation of proteomes routinely (Roche and Bruls, 2015).

Summary

Computational methods for prediction of protein structure are essential in the post-genomic era as experimental-based methods are unable to keep pace with the speed of genome sequencing. The production of 3D protein models, along with the structural annotation of entire proteomes, allows for both the interpretation of the proteins general function and the prediction of the binding site residues. These predictions may be exploited subsequently in in silico studies for the design of novel proteins for both medical and industrial applications, along with the development of drugs that will act as agonists or inhibitors for these proteins in order to modify their activity in disease pathways.

Future perspectives

With the advent of genomic medicine, where whole genome sequencing (WGS) is routinely undertaken on patients, to enable a better understanding of the patient's disease or illness, or to improve diagnosis. The key is to understand how the variant / mutation at the genome level affects the protein structure and function. Hence, it is important to examine the variant/mutation in the content of the protein, how this mutation will affect the protein structure and its function. As medical diagnostics needs to be carried out under time constraints, traditional methods for determining the structure of the mutant protein and how this mutation/mutations effect function is not possible, hence structure

prediction methods and structural annotation of proteomes are becoming more routinely used in medical diagnostics. In addition, as most diseases are multi-factorial, involving multiple proteins it will become extremely important to integrate the structural annotation of a patient's proteome with the whole genome sequencing studies. Furthermore, structurally annotation of proteomes can be useful in drug discovery pipelines and for numerous other industrial applications. In conclusion, we expect to see in the near future computational protein structure prediction and structural annotation of proteomes in personalised medicine, to enable more accurate disease diagnosis.

Cross-References

Alignment of Protein Sequences
Comparative Protein Structure Modeling
Domain Structure Classifications
Homology Modeling of Protein Structures
Macromolecular Crystallography: Overview
Protein NMR – Introduction
Protein Secondary Structure Prediction in 2012
Protein Structural Models – Evaluating Quality
Protein Structure Comparison Methods
Protein Structure Prediction
Proteins: Relationship Among Divergence of Sequence, Structure, and Function
Statistical Analysis and Prediction of Protein–Protein Interactions and Binding Sites

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
PubMed

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
PubMedCentral PubMed

de Lima Morais D, Fang H, Rackham OJL, Wilson D, Pethica R, Chothia C, Gough J. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucl Acids Res. 2011;39:D427–34.
PubMed

Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292:195–202.
PubMed

Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins. 2001;Suppl 5:127–32.
PubMed

Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K. Protein structure prediction center in CASP8. Proteins. 2009;77(Suppl 9):5–9.
PubMedCentral PubMed

Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., 2016. CASP11 statistics and the prediction center evaluation system. Proteins 84 Suppl 1, 15-19.

Lee J, Wu S, Zhang Y. Ab initio protein structure prediction. In: From protein structure to function with bioinformatics. London: Springer; 2009. p. 1–26.

Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227:1435–41.
PubMed

Maghrabi, A.H.A., McGuffin, L.J., 2017. ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. Nucleic Acids Res. 45, W416–W421.

McGuffin LJ. Aligning sequences to structures. In: Methods in molecular biology. Clifton: Humana Press; 2008a. p. 61–90.

McGuffin LJ. Protein fold recogntion and threading. In: Computational structural biology. London: World Scientific; 2008b. p. 37–60.

McGuffin LJ, Jones DT. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics. 2003;19:874–81.
PubMed

McGuffin LJ, Roche DB. Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics. 2010;26:182–8.
PubMed

McGuffin LJ, Street S, Sorensen SA, Jones DT. The genomic threading database. Bioinformatics. 2004;20:131–2.

PubMed

McGuffin LJ, Smith RT, Bryson K, Sorensen SA, Jones DT. High throughput profile-profile based fold recognition for the entire human proteome. BMC Bioinformatics. 2006;7:288.
PubMedCentral PubMed

McGuffin, L.J., Atkins, J.D., Salehe, B.R., Shuid, A.N., Roche, D.B., 2015. IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. Nucleic Acids Res 43, W169-173.

Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction – round VIII. Proteins. 2009;77(Suppl 9):1–4.
PubMed

Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.
PubMed

Roche DB, Buenavista MT, Tetchner SJ, McGuffin LJ. The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. Nucleic Acids Res. 2011;39:W171–6.
PubMedCentral PubMed

Roche, D.B., Bruls, T., 2015. The enzymatic nature of an anonymous protein sequence cannot reliably be inferred from superfamily level structural information alone. Protein Sci 24, 643-650.

Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195–7.
PubMed

Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21:951–60.
PubMed

Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C. Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res. 2008;36:D414–8.
PubMedCentral PubMed