PREDICTING PROTEIN STRUCTURES AND STRUCTURAL ANNOTATION OF PROTEOMES

Affiliation

Daniel Barry Roche
School of Biological Sciences,
University of Reading,
Whiteknights,
Reading, RG6 6AS,
UK

Maria Teresa Buenavista
School of Biological Sciences,
University of Reading,
Whiteknights,
Reading, RG6 6AS,
UK

Biocomputing,
MRC Harwell,
Harwell Science and Innovation Campus,
Oxfordshire OX11 0RD

Diamond Light Source,
Beamline B23,
Chilton,
Didcot OX11 ODE,
UK

Liam James McGuffin
School of Biological Sciences,
University of Reading,
Whiteknights,
Reading, RG6 6AS,
UK

Synonyms

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

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, is 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 currently <67,000 protein structures in the Protein Data Bank (PDB) and ~140 million sequences in the GenBank database (Figure 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.

Figure1_Growth_of_DB.tif. 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. B) As in A but enlarged to focus on the PDB entries and highlight the rate of increase in comparison to the sequence databases. C) The number of sequences in sequence databases and the number of protein structures in the PDB is plotted on a log scale against the years of entry.

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 (Figure 2) (Altschul et al., 1990; Altschul et al., 1997; Lipman and Pearson, 1985).

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 sequences 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 – **B**asic **L**ocal **A**lignment **S**earch **T**ool – (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 eleven 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.

Figure2_Multiple_seqeunce_alignment.tif. 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.

Critical Assessment of Techniques for Protein Structural Prediction (CASP)

The continual development of more advanced protein structure prediction tools is driven by the **C**ritical **A**ssessment of Techniques for Protein **S**tructural **P**rediction (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 six prediction categories: 1. Tertiary structure prediction –template based and free modelling, 2. Disorder prediction, 3. Domain prediction, 4. Contact prediction, 5. Quality assessment and 6. Binding site prediction (Moult et al., 2009).

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 modellers (Kryshtafovych et al., 2009). Server performance is extremely important, since they are the only choice for high throughput modelling. The number of servers involved in CASP has increased from 53 in CASP5 to 79 in CASP9, 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 modelling methods. If a structural template is available within the PDB, template based modelling methods such as homology modelling and fold recognition can be utilized, but if a structural template is unavailable free modelling will have to be utilized (Table 1).

Method category	**Requirements**	**Relative computational difficulty**	**Relative speed**	**Theoretical sequence coverage**
Homology / comparative modelling	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 modelling	The target sequence and/or a fragment library	Hard	Slow	Maximum

Table 1. Established techniques for the modelling 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).

Template based modelling

The success of template based modelling (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 modelling is divided into two subcategories: homology modelling and fold recognition. Homology modelling, also known as comparative modelling 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 modelling 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 made, then 3D models of the target structure can be built from the coordinates of the template structures (Figure 3) (Table 2).

A user friendly template based modelling prediction pipeline – IntFOLD-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) also integrates the ModFOLD 3.0 method, for model quality assessment, the DISOclust 2.0 method for protein intrinsic disorder prediction, the DomFOLD 2.0 method, for domain boundary prediction and the FunFOLD 1.0 method for the prediction of ligand binding site residues.

Figure3_TBM_pipeline.tif. Template based modelling pipeline, such as that used by IntFOLD-TS.

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/

Table 2. Some of the top publicly available template based modelling servers in CASP9.

Template free modelling

Free modelling has also been referred to as *ab initio* modelling, modelling from first principles or *de novo* modelling. Template free modelling is the prediction of a proteins tertiary structure from sequence without the use a protein structure as a template. The use of free modelling is necessary when a template cannot be found to predict the structure of the protein. Free modelling 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 modelling 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).

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 utilized 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 modelling is to utilize a compromised force field, with a large number of selected atom types. Within each atom type the physio-chemical 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 including hydrogen bonding and the local backbone stiffness of a polypeptide chain. The second energy term encompasses amino-acid or protein-sequence dependant information, which include: pairwise residue contact potentials, distance dependant 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 modelling. 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 modelling was first utilized in FRAGFOLD (Jones, 2001), and subsequently by the ROSETTA and the QUARK servers (Table 3).

Server	**URL**
chunk-TASSER	http://cssb.biology.gatech.edu/skolnick/webservice/chunk-TASSER/index.html
MULTICOM-NOVEL	http://casp.rnet.missouri.edu/nncon.html

QUARK	http://zhanglab.ccmb.med.umich.edu/QUARK
RAPTORX-FM	http://velociraptor.ttic.edu
Robetta	http://robetta.bakerlab.org/

Table 3. Some of the top publicly available free modelling web servers in CASP9.

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, 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) 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 hours 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).

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 sillico* 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.

Cross-References

→ …Nuclear Magnetic Resonance Spectroscopy
→ …X-ray diffraction and crystallography
→ …Domain structure family classifications
→ …Evolution of protein structures
→ …Homology modelling of protein structures
→ …Prediction of function from structure
→ …Structure comparison methods
→ …Applications of structure prediction methodology to biotechnology and medicine
→ …Biologically significant alignment of protein sequences
→ …Comparative modeling of protein structures
→ …Evaluating the quality of a protein structural model
→ …Fragment based methods for protein structure prediction
→ …The evolution of protein structures
→ …The relationship between the divergence of sequence and structure

Keyword

Protein structure prediction, sequence alignment, CASP, homology modelling, fold recognition, template free modelling, model quality assessment, structural annotation.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol *292*, 195-202.

Jones, D.T. (2001). Predicting novel protein folds by using FRAGFOLD. Proteins *Suppl 5*, 127-132.

Kryshtafovych, A., Krysko, O., Daniluk, P., Dmytriv, Z., and Fidelis, K. (2009). Protein structure prediction center in CASP8. Proteins *77 Suppl 9*, 5-9.

Lee, J., Wu, S., and Zhang, Y. (2009). Ab initio protein structure prediction. In From Protein Structure to function with Bioinformatics (London: Springer), pp. 1-26.

Lipman, D.J., and Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. Science *227*, 1435-1441.

McGuffin, L.J. (2008a). Aligning sequences to structures. In Methods in molecular biology (Clifton, N.J.) (Clifton, N.J.: Humana Press), pp. 61-90.

McGuffin, L.J. (2008b). Protein fold recogntion and threading. In Computational Structural Biology (London: World Scientific), pp. 37-60.

McGuffin, L.J., and Jones, D.T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics *19*, 874-881.

McGuffin, L.J., and Roche, D.B. (2010). Rapid model quality assessment for protein structure predictions using the comparison of multiple models without structural alignments. Bioinformatics *26*, 182-188.

McGuffin, L.J., Smith, R.T., Bryson, K., Sorensen, S.A., and Jones, D.T. (2006). High throughput profile-profile based fold recognition for the entire human proteome. BMC Bioinformatics *7*, 288.

McGuffin, L.J., Street, S., Sorensen, S.A., and Jones, D.T. (2004). The genomic threading database. Bioinformatics *20*, 131-132.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., and Tramontano, A. (2009). Critical assessment of methods of protein structure prediction - Round VIII. Proteins *77 Suppl 9*, 1-4.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

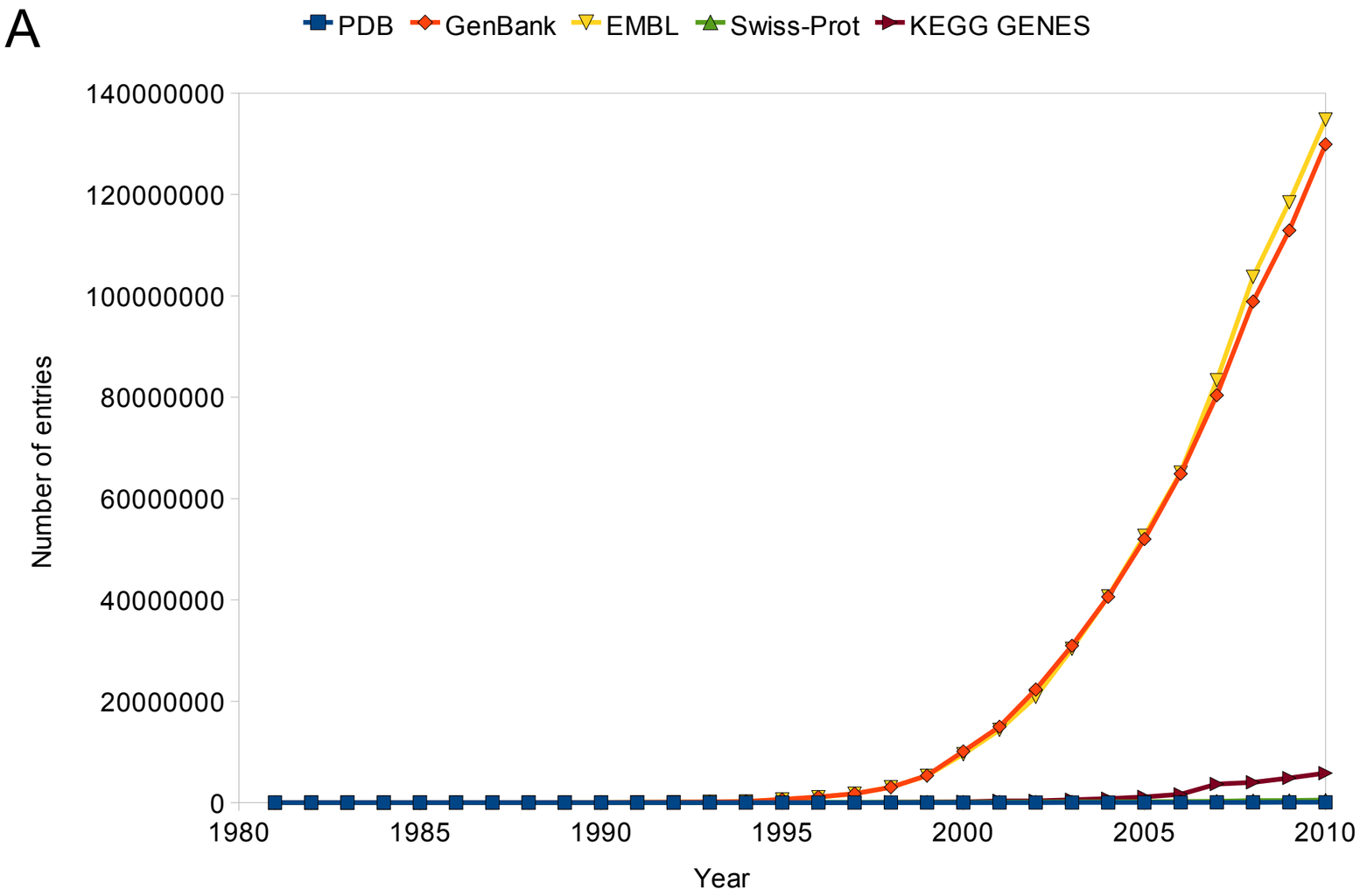
Roche, D.B., Buenavista, M.T., Tetchner, S.J., and McGuffin, L.J. (2011). The IntFOLD server: an integrated web resource for protein fold recognition, 3D model quality assessment, intrinsic disorder prediction, domain prediction and ligand binding site prediction. Nucleic Acids Res *In press*.

Smith, T.F., and Waterman, M.S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

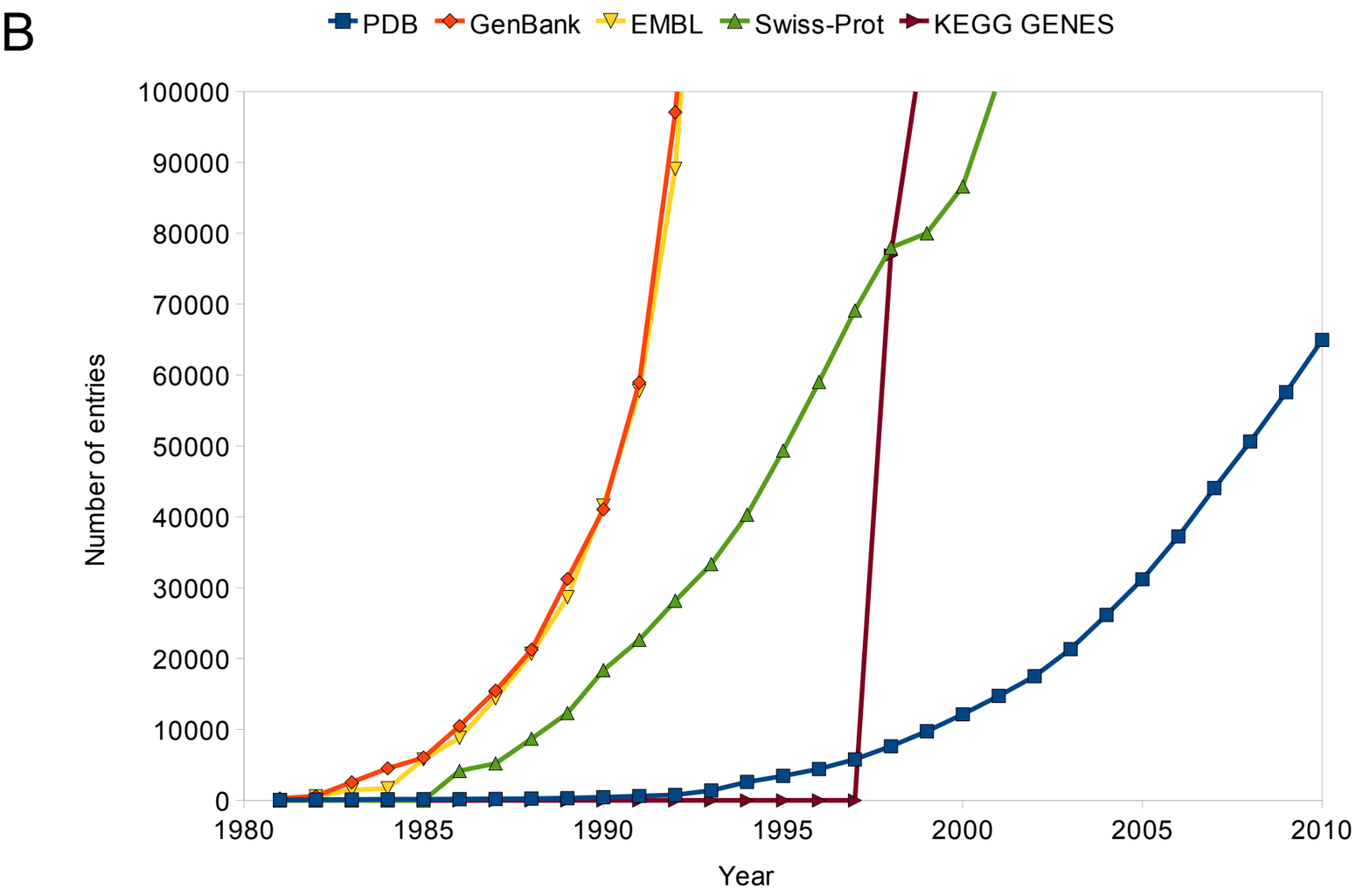
Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951-960.

Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X., and Orengo, C. (2008). Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res *36*, D414-418.

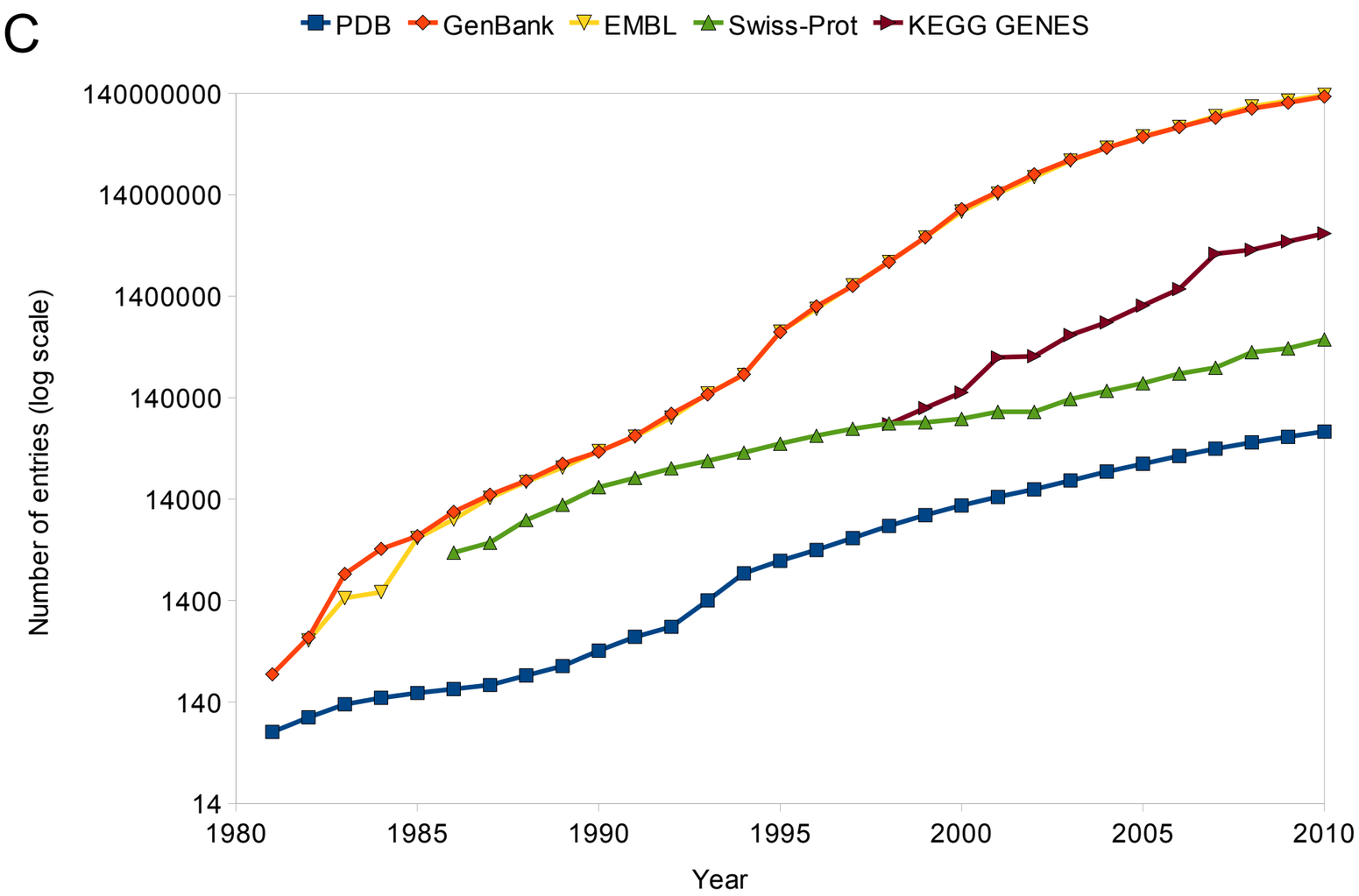
A

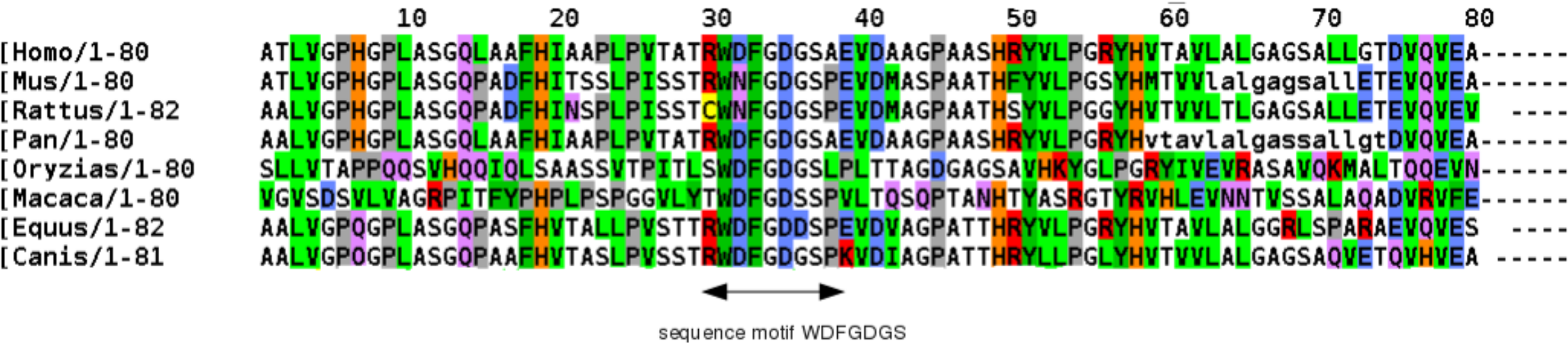

PDB
GenBank
EMBL
Swiss-Prot
KEGG GENES
140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
Number of entries
1980
1985
1990
1995
2000
2005
2010
Year

B


PDB
GenBank
EMBL
Swiss-Prot
KEGG GENES
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Number of entries
1980
1985
1990
1995
2000
2005
2010
Year

C


PDB
GenBank
EMBL
Swiss-Prot
KEGG GENES
140000000
14000000
1400000
140000
14000
1400
140
14
Number of entries (log scale)
1980
1985
1990
1995
2000
2005
2010
Year


10
20
30
40
50
60
70
80
[Homo/1-80 ATLVGPHGPLASGQLAAFHIAAPLPVTATRWDFGDGSAEVDAAGPAASHRYVLPGRYHVTAVLALGAGSALLGTDVQVEA------
[Mus/1-80 ATLVGPHGPLASGQPADFHITSSLPISSTRWNFGDGSPEVDMASPAATHFYVLPGSYHMTVVlalgagsallETEVQVEA------
[Rattus/1-82 AALVGPHGPLASGQPADFHINSPLPISSTCWNFGDGSPEVDMAGPAATHSYVLPGGYHVTVVLTLGAGSALLETEVQVEV ----
[Pan/1-80 AALVGPHGPLASGQLAAFHIAAPLPVTATRWDFGDGSAEVDAAGPAASHRYVLPGRYHvtavlalgassallgtDVQVEA------
[Oryzias/1-80 SLLVTAPPQQSVHQQIQLSAASSVTPITLSWDFGDGSLPLTTAGDGAGSAVHKYGLPGRYIVEVRASAVQKMALTQQEVN------
[Macaca/1-80 VGVSDSVLVAGRPITFYPHPLPSPGGVLYTWDFGDSSPVLTQSQPTANHTYASRGTYRVHLEVNNTVSSALAQADVRVFE------
[Equus/1-82 AALVGPQGPLASGQPASFHVTALLPVSTTRWDFGDDSPEVDVAGPATTHRYVLPGRYHVTAVLALGGRLSPARAEVQVES ----
[Canis/1-81 AALVGPQGPLASGQPAAFHVTASLPVSSTRWDFGDGSPKVDIAGPATTHRYLLPGLYHVTVVLALGAGSAQVETQVHVEA -----
sequence motif WDFGDGS

MANRLVSTASALVNANVSFECRLNFGTDVA
YLWNFGDDTIELGSSSSSHVYSREGEFTVEV
LARNNVSSTTLRKQLFIVREPCQPPPVKNM

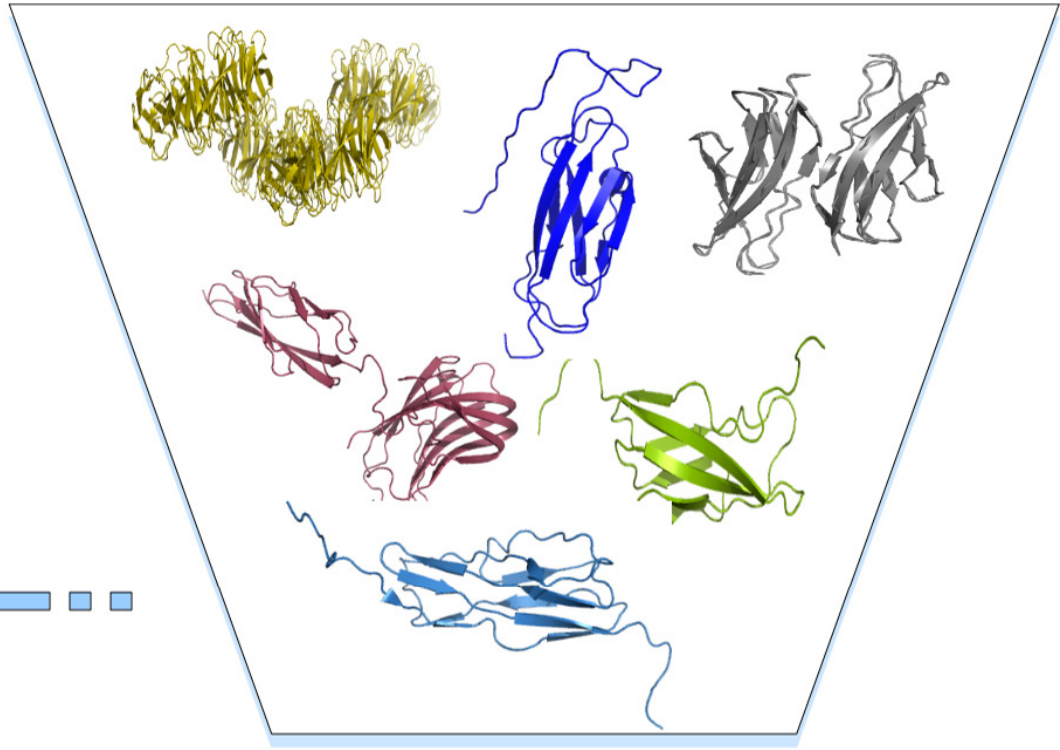
TARGET SEQUENCE

SEQUENCE – TEMPLATE ALIGNMENT

RANKING TEMPLATES

TOP TEMPLATE

STRUCTURAL REFINEMENT



TEMPLATE LIBRARY

TOP MODEL