universität wien

DISSERTATION

Titel der Dissertation

Likelihood of Protein Structure Determination

angestrebter akademischer Grad

Doktor/in der Naturwissenschaften (Dr. rer.nat.)

Verfasserin / Verfasser:	Suresh Kumar Sampathrajan
Matrikel-Nummer:	0648389
Dissertationsgebiet(lt. Studienblatt):	Molekulare Biologie (A 490)
Betreuerin / Betreuer:	Univ.-Prof. Dr. Kristina Djinovic

Wien, am 01. Dezember 2009

DEDICATION

To my Parents, brothers and friends; who have been constant source of inspiration in my life and career.

CONTENTS

1. TABLES AND FIGURES

2. Zusammenfassung

Strukturelle Genomanalyse (SG) beinhaltet die, mit hohem datendurchsatz verbundene bestimmung der dreidimensionalen struktur von makromolekülen durch experimentelle Methoden wie röntgenstrahlen-kristallographie und NMR spektroskopie. Eines der ziele von SG ist es, zeit und kosten der bestimmung von dreidimensionalen proteinstrukturen zu reduzieren, für die homologe strukturen noch nicht gelöst worden sind. Mehrere faktoren wie unregelmäßige conformationen, unzulässige selektion von domängrenzen und löslichkeit können die produktion von proteinkonstrukten für die strukturbiologie erschweren. Zuverlässige, auf aminosäuresequenz basierende prädiktoren zur berechnung von proteinkristallisation sind folglich von nöten.

Die vorhersage von unregelmäßigen konformationen ist essentiell, da diese schwierigkeiten in der kristallisation verursachen können. In dieser arbeit wird eine neue methode präsentiert, die es erlaubt, ungeordnete residuen auf basis der aminosäuresequenz mit hoher genauigkeit vorherzusagen, indem verschiedene, auf einer konsensusmethode basierende vorhersagemittel verwendet werden. Die Leistung dieser neuen methode ist signifikant besser als von jedem einzelnen, bisher erwähnten Prädiktor.

Zusätzlich ist es wichtig, die voraussetzungen für den quartärstatus eines proteins auf basis seiner sequenz vorherzusagen. Eine Proteinkette kann aus einem monomeren protein bestehen, oder kann, zusammen mit anderen ketten, oligomere komplexe formen, die entweder aus homo-oligomeren oder hetero-oligomeren bestehen können. Im letzten

fall muss vermieden werden, die dreidimensionale struktur eines einzelnen protomers zu bestimmen, weil es nicht funktionell ist und auch extrem schwer in löslicher form zu exprimieren ist. Es ist daher erstrebenswert, ein berechnungsmittel zu nützen, das vorherzusagen erlaubt, ob ein potentielles genprodukt teil eines permanenten und obligaten hetero-oligomeren komplexes ist. Hier wird eine neue, auf der aminosäuresequenz basierende methode präsentiert, um hetero-oligomere von monomer und homo-oligomeren proteinen und auch um monomere von homo-oligomeren mit hoher genauigkeit zu unterscheiden.

Das erfordernis von metallionen ist im design von strukturbiologischen experimenten ebenso wichtig. Metalloproteine bilden etwa ein drittel der proteoms. Die vorhersage von metalloproteinen hilft kristallographen, geeignetes wachstumsmedium für über-expressionsstudien auszuwählen und auch die wahrscheinlichkeit zu erhöhen, ein korrekt gefaltetes und funktionelles molekül zu erhalten. Hier wird gezeigt, dass die aufnahme von metallionen von proteinen auf basis der aminosäurenzusammensetzung und durch verwenden von lernfähigen analyseprogrammen mit hoher genauigkeit vorhergesagt werden kann.

Die ergebnisse in der vorliegenden Doktorarbeit stellen die basis für das sorgfältige design von Proteinkonstrukten dar. Diese computer basierenden selektionsmethoden sind hilfreich, um die auswahl von unmöglichen Zielen zu vermeiden – ein Muss in Strukturbiologie und Proteomics.

3. Summary

Structural Genomics (SG) involves the high-throughput determination of three-dimensional structures of macromolecules by experimental methods such as X-ray crystallography and NMR spectroscopy. One of the aims of SG is to reduce the time and cost in the determination of three-dimensional protein structures for which a homologous structure had not yet been solved. Several factors such as conformational disorder, improper selection of domain boundaries and solubility can hamper the production of protein constructs for structural biology. Reliable computational protein crystallization propensity predictors, based on amino acid sequences, are consequently required.

Prediction of protein conformational disorder is important since it can cause difficulty in crystallization. In this work, a new procedure is presented that allows one to predict disordered residues with high accuracy on the basis of amino acid sequences, by using a consensus method based on various prediction tools. The performance of this new procedure is significantly better than that of each individual predictor previously reported.

Furthermore, it is important to be able to predict the quaternary status requirements of a protein on the basis of its sequence. A protein chain can be a monomeric protein or it can form, together with other chains, oligomeric assemblies, which can be either homo-oligomers or hetero-oligomers. In the later case, it must be avoided to determine the three-dimensional structure of a single protomer, since it will not be functional and it will also be extremely difficult to express in a soluble form. It is thus desirable to have a

computational tool that allows one to predict if a potential gene product is a part of permanent and obligate hetero-oligomeric assembly. A new method is presented for discriminating hetero-oligomers from monomeric and homo-oligomeric proteins and also between monomers and homo-oliogmers with high accuracy on the basis of amino acid sequences.

Metal ion requirements are also important in designing structural biology experiments. Metalloproteins constitute about one-third of the proteome. Prediction of metalloprotein helps crystallographers to select the proper growth medium for over-expression studies and also to increase the probability of obtaining a properly folded and functional molecule. Here it is shown that the uptake of metal ions by proteins can be predicted with high accuracy on the basis of the amino acid composition and by using machine learning methods.

The results described in the present Thesis provide a basis for the careful design of protein constructs. These computational screening methods are helpful to avoid the selection of 'impossible' targets- a must in structural biology and proteomics.

4. Introduction

4.1. Structural genomics

Structural genomics (SG) aims to determine the three-dimensional shapes of all important biological macromolecules, with primary focus on proteins. The main goal of the project is to expand the structural knowledge of biological macromolecules, and lowering the average cost of structure determination through high-throughput methods (Joachimiak 2009).

Structural genomics has now become a driving force behind new developments in protein structure prediction technology, aiming to automate, and consequently expedite, all areas of the experimental pipeline, ultimately benefiting the structural biology community as a whole. Recent analyses of structures released by the initiatives have highlighted the significant contribution they are now making in both the scope and depth of our structural knowledge of protein families, especially when compared to the relative contribution of non-structural genomics structures (Marsden et al. 2007).

4.1.1. Protein Structure Initiative and other SG initiatives

Protein Structure Initiative (PSI) (Hendrickson 2007) is one among the structural genomics projects, which aim to determine three-dimensional protein structures, for which a homologous experimental structure had not yet been solved.

Table 1. Some other SG Initiatives, besides the PSI.

Group or SG center	*Key ideas*	*Web address*
Berkeley Structural Genomics Center (BSGC)	To obtain a structural complement of two minimal genomes, *Mycooplasma genitalium* and *Mycooplasma pneumoniae*, two related human and animal pathogens	http://www.strgen.org/
Protein Structure Factory (PSF)	Technology development; human proteins	http://www.proteinstrukturfabrik.de/
Centre for structural Genomics of Infectious Diseases (CDGID)	To use high-throughput (HTP) structural biology technologies to experimentally characterize the three-dimensional structure of targeted proteins from major human pathogens	http://www.csgid.org/csgid/cake/
Oxford Protein Production Group (SPINE)	To determine structures of proteins and protein complexes from bacteria and human, viral pathogens	http://www.spineurope.org/
RIKEN Structural Genomics	To determine the 3D structures of human, mouse, bacteria, and archaea	http://www.riken.go.jp/
TB Structural Genomics Consortium (TB)	Determination and analysis of protein structure from *Mycobacterium tuberculosis* proteins, and genomics consortium large-scale collaboration	http://www.doembi.ucla.edu/TB/
Vizier project	Identification of potential new drug targets against RNA viruses through comprehensive structural characterization	http://www.vizier-europe.org/
The structural Genomics Consortium, Toronto	To determine the 3D structures of human proteins of therapeutic relevance to diseases such as cancer and diabetes metabolic disorders	http://www.sgc.utoronto.ca
Targeted Proteins Research Program (TPRP)	The program aims to reveal the structure and function of proteins that have great importance in both academic research and industrial application.	http://www.tanpaku.org

The long-range goal of the PSI is to make the three-dimensional structures of most proteins easily obtainable on the basis of their corresponding Deoxyribonucleic acid (DNA) sequences (Norvell and Berg 2007). Structural Genomics centers contribute about half of new structurally characterized families of proteins, and PSI centers account for about two-thirds of the worldwide SG output.

In the first phase of the Protein Structure Initiative, major goals were to lower the cost and to increase the success rates of structure determination by developing new methodologies to construct and automate the protein production (Blundell 2007; Norvell and Berg 2007). In the second phase of the Protein Structure Initiative, the major goals were focused on the structural coverage of sequence families of biological importance, beside the development of new methodology for challenging classes of protein (i.e., integral membrane proteins and protein-protein complexes) (Matthews 2007; Dessailly et al. 2009). Beside Protein Structure Initiative, several other SG Initiatives are in progress and some of them are shown in table 1

4.1.2. Target selection

In the Protein Structure Initiative, targets of structure determination are chosen from large protein sequence families for which there is no structural information and from very large phylogenetically diverse protein families, which are inadequately characterized at the level of the three-dimensional structure. Once approved from a scientific committee, each PSI production center selects non-redundant target protein families. After a protein family is selected, individual targets are chosen at the center (Burley et al. 2008).

4.2. Experimental pipeline of protein crystal structure determination

The determination of a protein structure by experimental X-ray crystallographic analysis involves the following steps.

4.2.1. Expression and purification of proteins

For high-throughput analysis, milligram quantities of very pure and homogeneous protein are usually required for successful crystal growth and crystal structure determination. Over-expression in bacteria or in another suitable system is consequently a necessity in most cases. The majority of structural genomics consortia are pursuing high-throughput protein expression through constructs expressed in *Escherichia coli*.

Purification of over expressed protein is greatly simplified for high-throughput studies through the use of constructs in which the target gene is fused to an affinity tag, whereby the tag can be placed at either the amino- or the carboxy-terminal end of the target protein, with a number of options in construct design. Polyhistidine-tag (His-tag) is a widely employed method.

4.2.2. Protein crystallization

To obtain well diffracting, well-ordered protein single-crystals is the vital aim of protein crystallography. Protein crystals can be obtained by slowly withdrawing solvent from a highly concentrated protein solution.

In the most common methods of growing protein crystals, purified protein is dissolved in an aqueous buffer containing a precipitant such as ammonium sulfate or polyethylene glycol, at a concentration just below that necessary to precipitate the protein. Then water is removed by controlled evaporation to produce precipitating conditions, which are maintained until crystal growth ceases.

There are many methods used for protein crystallization. The most commonly used is vapour diffusion. Vapour diffusion has two variants known as the hanging drop and sitting drop methods (see figure 1) (McPherson 2004).

A few micro liters solution of purified protein is mixed with an equal amount of the reservoir solution, giving precipitant concentration about 50% of that required for protein crystallization. This solution is suspended as a droplet underneath a cover glass, which is sealed onto the top of the reservoir grease. Because the precipitant is the major solute present, vapour diffusion in this closed system results in net transfer of water from the protein solution to the reservoir, until the precipitant concentration is the same in both solutions. The reservoir is much larger than the protein solution; the final concentration of the precipitant in the protein solution is nearly equal to that in the reservoir. When the system comes to equilibrium, net transfer of water ceases, and the protein solution is maintained at the optimal precipitant concentration.

The hanging drop method differs from the sitting drop method in the vertical orientation of the protein solution drop within the system. It is important that both the hanging drop

and the sitting drop methods require a closed system, that is, the system must be sealed off from the outside using an airtight container or high-vacuum grease between glass surfaces.

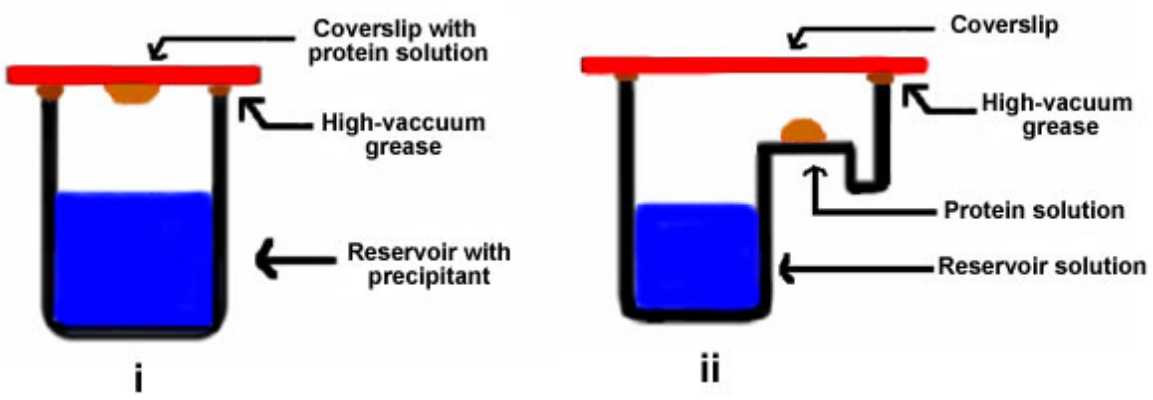


Figure 1. Principal techniques of protein crystallization. **(i)** Hanging drop method (ii) Sitting drop method (modified from: http://www.protocol-online.org).

4.2.3. Structure determination

Once suitable protein crystals become available, X-ray diffraction data are collected. Diffraction data are measured using monochromatic X-ray from a sealed tube generator, a rotating anode X-ray generator or from a synchrotron source. It is obviously important to determine the phases of the diffraction data. They can be determined by two methods: molecular replacement and isomorphous replacement. With an initial set of phases one can calculate an approximate three-dimensional density map of the protein structure. Model refinement is carried out against the experimentally measured diffraction data and may include the addition of well-ordered solvent molecules.

4.3. Crystallization propensity predictors and databases

Several crystallization propensity predictors were developed for predicting if a certain sequence will be suitable for expression, purification and crystallization. Some predictors were focused on specific problems in the experimental pipeline like predicting conformational protein disorder, identification of protein domain boundaries and prediction of post-translation modifications, as these factors may hinder in a successful structure determination. It is clear that the three-dimensional structure of a conformationally disordered protein cannot be determined experimentally. Several predictors were based on amino acid sequence. Although none of the available methods for disorder prediction are fully reliable on its own, it is often necessary to consider merits and demerits and to combine them to achieve reliable prediction. The identification of the domain boundaries is often a crucial problem during structural biology experiments, in selecting and fine tuning of the amino acid construct. In general it is major concern problem with regard to large proteins, usually composed of several separate structural domains. Prediction of post-translational modifications and translocation signals are important because they are often related to transitions between an ordered and a disordered conformational state of the protein and also to the expression system (Carugo et.al.2007a).

These predictors help to avoid time consuming expensive experiments on 'impossible targets'. These approaches may be valuable in structural genomic projects when there is a desire to rank targets according to likely success or for structural biologists to handle

their specific problems and optimize their experimental strategy (Carugo et.al.2007b). There are several crystallization propensity predictors as follows.

4.3.1. SECRET

A sequence-based crystallizability evaluator (Smialowski et al. 2006) is a machine-learning approach that predicts protein crystallizability to support target selection process in structural genomics. It uses a meta-method as a classification algorithm, consisting of two layered structures with support vector machines as a primary classifier, and a Naive Bayes as a second level classifier. It is available online at http://mips.helmholtz-muenchen.de/secret/secret.seam.

4.3.2. CrystalP

CrystalP (Chen et al. 2007) uses a novel feature-based sequence representation which is based on frequency of collocated amino acid pairs in the sequence and applies a Naive Bayes classifier. This method can be used to predict if a small and medium size (<200 amino acids), proteins can be crystallized. Additionally, features used by crystalP may help to discover intra-molecular markers that influence protein crystallization. Unfortunately it is not available as a web-server or as stand alone software.

4.3.3. OBscore

The OB-score (Overton and Barton 2006) estimates a protein propensity to produce diffraction-quality crystals, on the basis of calculated isoelectric points and hydrophobicity values. High positive OB-score may be used to indicate the proteins that should be more prone to structure determination while low OB-score can suggest more challenging proteins. The OB-score software is freely available from http://www.compbio.dundee.ac.uk/obscore.

4.3.4. Taro

TarO (Overton et al. 2008a) is a predictor focused on structural genomics target selection/optimization. It includes crystallization propensity predictions, orthologue searching, and many other sequence-based calculations. The results are available through an annotated multiple sequence alignment. It is available as a guest account at http://www.compbio.dundee.ac.uk/taro/.

4.3.5. ParCrys

ParCrys (Overton et al. 2008b) implements a parzen window approach based on the calculated isoelectric point, hydrophobicity and the frequencies of S,C,G,F,Y,M residues. It uses the Protein Data Bank (PDB) as the training data. It is available online at http://www.compbio.dundee.ac.uk/parcrys.

4.3.6. TargetDB

TargetDB (Chen et al. 2004) is a target registration database that provides information on the experimental progress and status of targets selected for structure determination, which includes protein target data from the NIH structural genomics centers and a number of international structural genomic initiatives.

A number of other worldwide structural genomics centers have also contributed data to TargetDB on a voluntary basis. TargetDB, which is hosted by the Protein Data Bank (RCSB PDB), provides status information on target sequences and tracks their progress through the various stages of protein production and structure determination. TargetDB includes information about status category as described in table 2.

Table 2. Various steps along the structural biology pipeline monitored by the TargetDB database.

Category	***Status***
Target preparation	Cloned, expressed. soluble, purified
Crystallization	Crystallized, diffraction-quality crystals, diffraction, crystal structure
NMR structure determination	Heteronuclear Singe Quantum Coherence (HSQC), NMR assigned, NMR structure
Deposition status	In PDB
Work stopped	
Test target	

TargetDB has the query capabilities and one can search the database by sequence search method, based on project site, target ID, protein name, source organism, date of last modification, and the current status of the target. The status search category is also available with option to indicate if work has been stopped on a particular target. Based on

status search information for each target, two types of summary reports are generated. One gives the progress for an individual target (or lists of targets) according to its change in status over time and other describes the aggregate status information of each structural genomics center. TargetDB is available at http://targetdb.pdb.org/

4.4. Machine learning

By dictionary definition, machine learning is defined as "to gain knowledge, or understanding of, or skill in, by study, instruction, or experience," and "modification of a behavioral tendency by experience". But broadly it can be defined that a machine learns from its inputs or in response to external information and it changes its structure, program, or data in such a way that it is expected to improve its performance in future.

Machine learning is like programming computers to optimize a performance criterion using novel data or past experience. It consists of modelling with defined parameters and training data. Learning is the execution of the program to optimize the parameters of the model using training data or past experience. In machine learning, first, we need efficient algorithms to solve the optimization problem, as well as to store and process the massive amount of training data. Second, once a model is learnt, its representation and algorithmic solution for inference needs to be efficient, in space and time complexity.

4.4.1. Supervised learning

Supervised learning is a machine learning technique used for prediction of the value of the function for any valid training data containing input objects. The goal of supervised classification is to find a functional mapping between the input data X, describing the input pattern, to a class label Y (e.g. -1 or +1), such that Y= f (X). The construction of the mapping is based on the so-called training data, supplied to the classification algorithm. Curve-fitting is a simple example of supervised learning of a function (Larranaga et al. 2006).

4.4.2. Unsupervised learning

Unsupervised learning is a type of machine learning technique used to determine how the data are organized. It differs from the supervised learning since the training vectors lack function values (Goldbaum 2005). Principal component analysis and cluster analysis are examples of unsupervised learning/pattern recognition.

4.4.3. Semi-supervised learning

Semi-supervised learning is a machine learning technique intermediate between unsupervised learning and supervised learning (Ernst et al. 2008). The semi-supervised technique makes use of both labeled and unlabeled data for training, sometime with considerable improvement in learning accuracy.

As the supervised learning technique use only labeled data to train, sometime it is very time consuming, expensive and also difficult and it requires lot of efforts of experienced human annotators. On the contrary, the absence of annotations makes unsupervised methods relatively easy. However, their results are often ambiguous. Semi-supervised learning solves this problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers (Kall et al. 2007). Semi-supervised has advantages like less human effort and also can give higher accuracy.

4.4.4. Commonly used classifiers

A classifier is a function that takes the values of various features (independent variables) and predicts the class (dependent variable) (Pereira et al. 2009). The most commonly used classifiers are briefly described in the next few pages.

4.4.4.1. K-nearest neighbor classification

K-nearest neighbor is the method of instance based learning. Each new instance (the query) is compared with existing ones using a distance metrics, and the closest existing instance is used to assign the class to the new one. Sometime, more than one nearest neighbor is used, and the majority class of the closest k neighbors is assigned to the new instance. This is the generic k-nearest-neighbor method. This approach is highly intuitive and gives low classification errors, but it is computationally expensive and requires a large memory to store the annotated data (Gil-Pita and Yao 2008).

4.4.4.2. Naive Bayes

Naive Bayes is a simple probabilistic classifier based on Bayes rules and "naively" assumes independence of the events. The assumption that the variables are independent can harm the performances, since it is obvious that the conditional independence assumption is rarely true in most real-world applications. An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification (Yan et al. 2006).

4.4.4.3. Decision trees

Decision tree algorithms solve the classification problem by repeatedly partitioning the input space, so as to build a tree whose nodes are as pure as possible. Nodes of a decision tree involve the testing of a particular attribute. Usually, the test compares an attribute value with a constant. Classification of a new instance is achieved by moving it from top to bottom along the branches of the tree, starting from the root node, until a terminal node is reached (Tan et al. 2005). Decision trees are really simple and effective for small datasets, but for large datasets they require large storage memory (Salzberg 1995).

4.4.4.4. Neural networks

Neural networks are a computational model inspired by the connectivity of neurons in animate nervous systems.

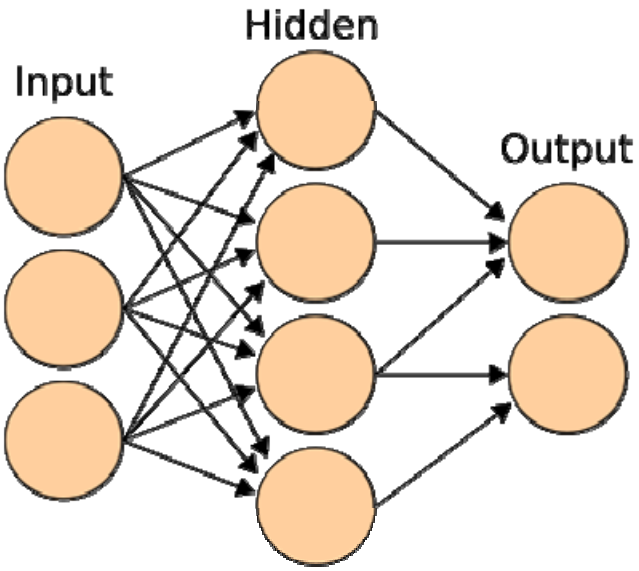


Figure 2. Schematic representation of a neural network. Each circle in the hidden and output layer is a computational element known as a neuron.

In the figure 2, each circle denotes a computational element referred to as a neuron, which computes a weighted sum of its inputs. If certain classes of nonlinear functions are used, the function computed by the network can approximate any function (Melville et al. 2009).

4.4.4.5. Support Vector Machines

Support Vector Machines (SVMs) are machine learning technique related to supervised learning methods used for classification and regression.

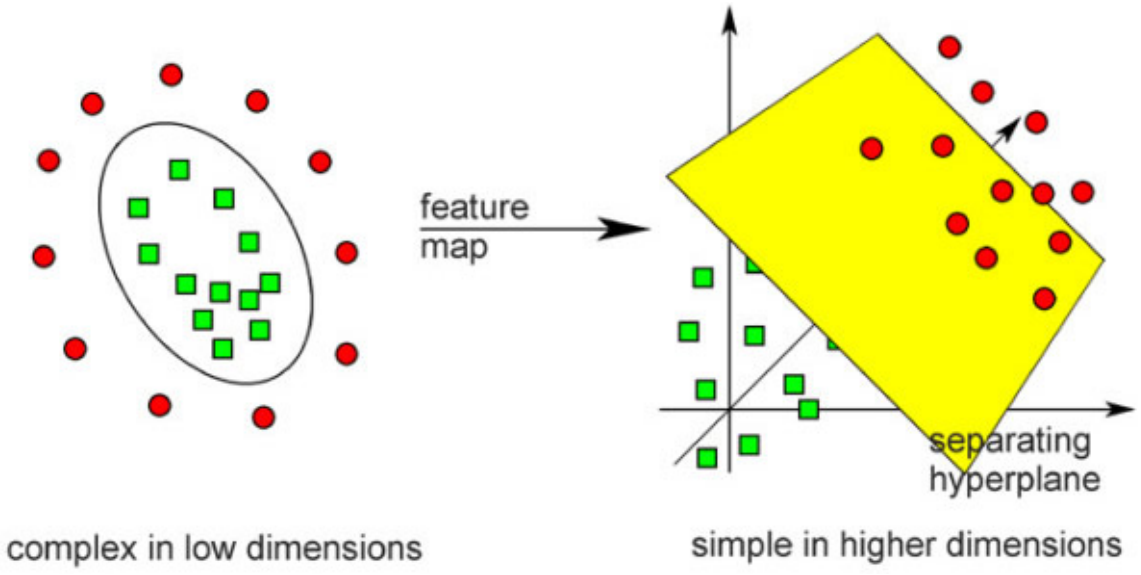


Figure 3. Schematic diagram of support vector machine (modified from Byvatov and Schneider 2003). By using a kernel, the original bi-dimensional space is transformed into the feature tri-dimensional space, on which the instances can be optimally separated by the yellow plane.

Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear discriminative function that separates them as widely as possible. The SVM finds a large margin separation between the training examples (figure 3). Hence, the large margin then ensures that new instances can be correctly classified as well (Byvatov and Schneider 2003). Kernels are often used to linearize and simplify the construction of an optimal separation between the learning sets.

4.4.5. Validation

Validation is a method to assess how model is likely to fit new data. Validating the performance of predictive models is the single most important step. The different approaches of data validation are as follows.

4.4.5.1. Cross-validation

Cross-validation is the most widely used method for estimating prediction reliability. To determine the accuracy of the prediction, the predicted values are compared with actual values obtained from a new sample of subjects or with values from a sample obtained at the time of the original data collection but held out from the initial analysis. One fold of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called test set).

4.4.5.2. K-fold cross-validation

K-fold cross validation splits the data into k approximately equally sized partitions. The induction algorithm is executed k times; each time it is trained on k-1 partitions and the generated hypothesis is tested on the rest of the data, which serves as a test set.

4.4.5.3. Jackknife

Jackknife involves in systematically recomputing the estimate leaving out one observation at a time from the sample set. Jackknifing is similar to bootstrapping and may in many situations yield similar results. However, the jackknife is easier to apply to complex sampling schemes, such as multi-stage sampling with varying sampling weights, than the bootstrap.

4.4.5.4. Confusion matrix and figures of merit

A confusion matrix contains information about actual and predicted classifications done by a classifier. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. An example for confusion matrix for a two class classifier is shown in table 3.

The entries in the confusion matrix denotes as follows: **a** is the number of correct predictions that an instance is true negative, **b** is the number of incorrect predictions that an instance is false positive, **c** is the number of incorrect predictions that an instance is false negative, and **d** is the number of correct predictions that an instance is true positive.

Table 3. Example of Confusion matrix.

Actual	***Predicted***	
	Negative	Positive
Negative	**a**	**b**
Positive	**c**	**d**

Several standard terms have been defined for the two class matrix. True Positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated TP=d/c+d. The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated FP=b/a+b. The true negative rate (TN) is defined as the proportion of negatives cases that were classified correctly, as calculated TN=a/a+b. The false negative rate (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated FN=c/c+d.

Alternative definitions are also commonly used, like sensitivity, specificity, Matthews correlation coefficient, accuracy, precision and probability excess. Sensitivity measures the proportion of actual positives which are correctly identified. Specificity measures the proportion of negatives which are correctly identified. The Matthews correlation coefficient (MCC) takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The accuracy is the proportion of true results (both true positives and true negatives) in the population. On the other hand, precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives). They are calculated as follows.

$$Sensitivity = \frac{d}{d+c}$$

$$Specificity = \frac{a}{a+b}$$

$$Accuracy = \frac{d+a}{a+b+c+d}$$

$$Precision = \frac{d}{d+b}$$

$$Matthews\_cc = \frac{(a \,.\, d - b \,.\, c)}{\sqrt{(b+d)(c+d)(a+b)(a+c)}}$$

$$Probability\_excess = Sensitivity + Specificity - 1$$

While sensitivity, specificity, precision and accuracy may be misleading for unbalanced data sets. [(a+b) >> (c+d) or (a+b) << (c+d)], the Matthews correlation coefficient (MCC) and the probability excess are more robust and tolerate unbalanced samplings. The sensitivity, specificity and accuracy values may range, by definition, from 0 to 1, all the other figures of merit can range between -1 and 1.The figures of merit which has higher values are associated with more reliable predictions, in particular positive values of the Matthews correlation coefficient and of the probability excess are associated with good and non-random predictions.

4.5. Protein conformational disorder prediction

4.5.1. Introduction

Soluble proteins may consist of globular units and non-globular units. Globular units are composed of regular secondary structure elements whereas the non-globular units are composed of disordered, unstructured, and flexible regions without regular secondary structure elements (Linding et al. 2003a). Recently it is well known that many functionally important protein segments occur outside of globular units (Wright and Dyson 1999; Dunker et al. 1998). Intrinsically Disordered Proteins (IDPs) (also known as intrinsically unstructured, unfolded, or rheomorphic proteins) lack a well defined 3D structure and exhibit a multitude of conformations that dynamically change over time and population (Dunker et al. 2008; Meszaros et al. 2009).

Many proteins cannot be over-expressed, purified or crystallized. Native conformational disorder of proteins is one of the main obstacles facing structural biology analyses. Moreover, in structural genomics initiatives, it is becoming increasingly important to identify the intrinsic protein disorder during the target selection process (Smialowski et al. 2006; Ferron et al. 2006).

4.5.2. Natively disordered proteins

No commonly agreed definition of protein disorder exists. "Natively disordered or unfolded proteins are proteins that do not form a stable three-dimensional structure in their native state. A disordered protein can be either completely unfolded or comprise both folded and unfolded segments" (Fink 2005). While completely unfolded proteins

carry out their function by means of regions that lack specific 3D structure and exists as an ensembles of flexible molecules, in partially unfolded proteins only localized regions lack organized structure.

4.5.3. Biological importance of protein disorder

The native unstructured proteins are found to participate in many biological processes and commonly occur in cell signaling pathways, DNA transcription and replication and protein translation (Vucetic et al. 2007). Protein disorder is important for understanding protein function as well as protein folding pathways (Uversky 2002; Tompa 2002). More than 180 such proteins are known including Tau, Prions, Bcl-2, p53, 4E-BP1, elF1A and HMG proteins (Linding et al. 2003a). They are thought to become ordered when they are bound to another molecule (e.g. CREB-CBP complex) although little is understood about the cellular and structural meaning of disorder. Due to the abnormal aggregation patterns of these proteins, they are involved in major protein diseases such as Parkinson's and Alzheimer's syndromes (Forloni et al. 2002).

4.5.4. Experimental determination of intrinsic disorder

Protein disorder is indirectly observed with a variety of experimental methods, such as X-ray crystallography, NMR, Raman spectroscopy, CD spectroscopy, and hydrodynamics measurements.

4.5.4.1. X-ray crystallography

Disorder leads to missing electron density in protein structures determined by X-ray crystallography. Two types of disorder have been recognized: static and dynamic (Huber 1979). Disorder is static when different molecules are rigid and adopt different conformations. It is, on the contrary, dynamic when molecules are flexible and oscillate between various conformations. If a region exists as an ensemble of φ and ψ angles, whether static or dynamic, it is intrinsically disordered. The major disadvantage of X-ray crystallography is that it requires additional experimental confirmation whether the missing electron density is a wobbly domain, is intrinsically disordered, or is the result of technical difficulties (Dunker et al. 2001).

4.5.4.2. Nuclear Magnetic Resonance spectroscopy (NMR)

Protein three-dimensional structures can be determined in solution by NMR. Under favorable circumstances, NMR provides motional information on a residue-by-residue basis by means of a variety of different isotopic labeling and pulse sequence experiments (Dunker et al. 2001).

4.5.4.3. Circular Dichroism (CD) spectroscopy

Structural information for proteins in solution is also provided by circular dichroism. The circular dichroism spectroscopy can give semi-quantitative information by combined use of near and far UV CD. It does not provide clear information for the proteins that contain both ordered and disordered regions (Dunker et al. 2001).

4.5.4.4. Protease digestion

Protease digestion is a well recognized method, which gives insight into protein structure and flexibility. The protein digestion method is particularly useful when used in combination with other methods. Protein digestion along with X-diffraction method helps to sort out whether a region of missing electron density is due to a wobbly domain or to intrinsic disorder. Protein digestion is useful when coupled with CD spectra, which lack position-specific information. Finally, the combination of proteolysis and mass spectrometry for fragment identification can indicate the presence of intrinsically disordered regions (Dunker et al. 2001).

4.5.4.5. Stokes radius determination

Random coil disorder has also been detected by various methods for obtaining stokes radius such as small-angle X-ray scattering or size exclusion chromatography (Dunker et al. 2001).

4.5.5. Protein conformational disorder predictors

Various computational predictors were developed for predicting protein conformational disorder. Since there is no unique definition of protein disorder, each of the predictors has its own definition and algorithms. Some of them have different versions of the same basic algorithm. Some of the predictors are based on datasets of ordered/disordered proteins. Others are based on physicochemical trends and observations. All the predictors mentioned here are available as web servers (table 4).

.

Table 4. Summary of the web servers available for the prediction of protein conformational disorder.

Server name	*What is predicted?*	*Web address*	*Reference*
Disembl_hotloops	Loops with high B-factors (highly mobile loops) from X-ray crystal structures using neural networks	http://dis.embl.de/	(Linding et al. 2003a)
Disembl_loops	Residues within loops/coils (regions devoid of regular secondary structure) using neural networks	http://dis.embl.de/	(Linding et al. 2003a)
Disembl_remark465	Defined by the REMARK465 (regions lacking electron density) lines in the PDB files using neural networks	http://dis.embl.de/	(Linding et al. 2003a)
Disopred	Predicts residues with missing atomic co-ordinates in the PDB files	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html	(Ward et al. 2004)
Drip-pred	Secondary structure based using Kohonen's self-organizing maps	http://www.sbc.su.se/~maccallr/disorder/	unpublished
Foldindex	Regions that have a low hydrophobicity and high net charge (either loops or unstructred regions)	http://bip.weizmann.ac.il/fldbin/findex	(Prilusky et al. 2005)
Globplot_B	Uses a propensity scale called "Russel/Linding" based on the hypothesis that the tendency for disorder for a given aminoacid to be in to be either in regular secondary structures (α-helices or β-strands) or outside of them ('random coil', loops, turns etc.).	http://globplot.embl.de/	(Linding et al. 2003b)
Globplot_R	Uses propensity scale based on missing coordinates in X-ray structures as defined by the REMARK465 lines in the PDB files.	http://globplot.embl.de/	(Linding et al. 2003b)
Iupred_L	Predictions are based on the algorithm that evaluates the energy of inter residues interaction. Residues that do not have the capacity to form sufficient inter-residue interactions predicted to be disordered	http://iupred.enzim.hu/index.html	(Dosztanyi et al. 2005)
Iupred_S	Missing residues in X-ray structure as defined by the REMARK465 lines in the PDB files.	http://iupred.enzim.hu/index.html	(Dosztanyi et al. 2005)
Prelink	Predictions are based on amino acid composition and on hydrophobic cluster content	http://genomics.eu.org/spip/PreLink	(Coeytaux and Poupon 2005)
Ronn	Uses bio-basis function neural network pattern recognition algorithm to the detection of natively disordered (lack well-defined 3D structure) regions in proteins.	http://www.strubi.ox.ac.uk/RONN	(Yang et al. 2005)

4.5.6. Protein trinity/quartet hypothesis

There are two major views of categorization of the form of IDPs (figure 4). According to Dunker and Obaradovic (Dunker and Obradovic 2001), IDPs exist in two different forms: molten globule (collapsed) and random coil-like.

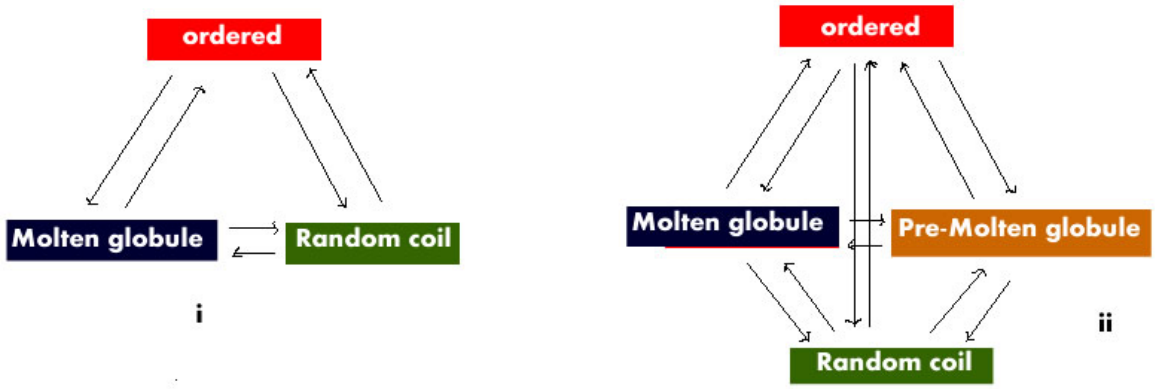


Figure 4. (i) The Protein Trinity Hypothesis (ii) The Protein quartet hypothesis.

On the contrary, Uversky suggested the existence of another form called pre-molten globule (Uversky 2002), which is intermediate between the fully extended and the molten-globular conformations. Along with the folded conformation, they form the protein trinity (Dunker and Obradovic 2001) or the protein-quartet (Uversky 2002) hypothesis.

4.5.7. Database of protein disorder

(i) ProDDO

The first public database of disordered proteins regions was limited to the PDB entries only. It does not provide information about the type of disorder or the function of disordered regions (Sim et al. 2001). The database is now no longer maintained.

(ii) Disprot

Disprot overcomes the above limitations. It is a curated database that provides information about proteins that lack fixed 3D structure in their putatively native states, either entirely or in part. The database contains experimentally characterized IDPs and includes functional information for many of the IDPs and regions (Sickmeier et al. 2007). In its first public release of February 2004, DisProt contained 154 proteins (190 disordered regions); whereas in June 2009 the database contained 523 proteins (1195 disordered regions). The database can be accessed at http://www.disport.org.

4.5.8. Protein disorder and the Protein Data Bank

The Protein Data Bank (Bernstein et al. 1977; Berman et al. 2000) stores atomic coordinates issued from X-ray and NMR studies. Some of the PDB entries contain the lines labeled with "REMARK 465", which list the residues that were not detectable experimentally. In a recent survey of the PDB, limited to the crystal structures (Kumar and Carugo 2008), it is observed that a considerable number of structures have conformational disordered. In 22% of them, more than 5% of the residues are disordered.

However, only about 2% of the crystal structures contain more than 20% of the residues that lack a well defined structure as shown in figure 5.

It was also observed that resolution tends to decrease if the amount of disorder increases as shown in figure 6 although the resolution decrease is not as large as it might be expected.

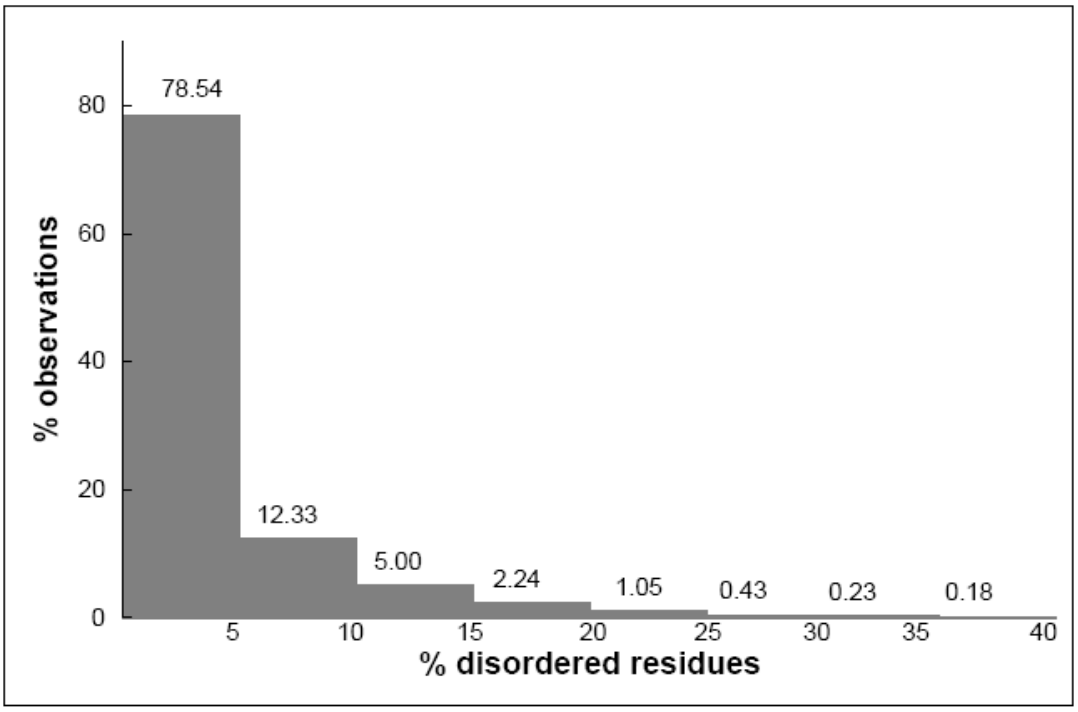


Figure 5. Distribution of protein crystal structures as a function of the percentage of disordered residues they contain. The data were taken from the Protein Data Bank; a residue was considered to be disordered if not observed in the crystallographic electron density maps; the total number of residues was taken from the SEQRES record of the PDB files (Kumar and Carugo 2008).

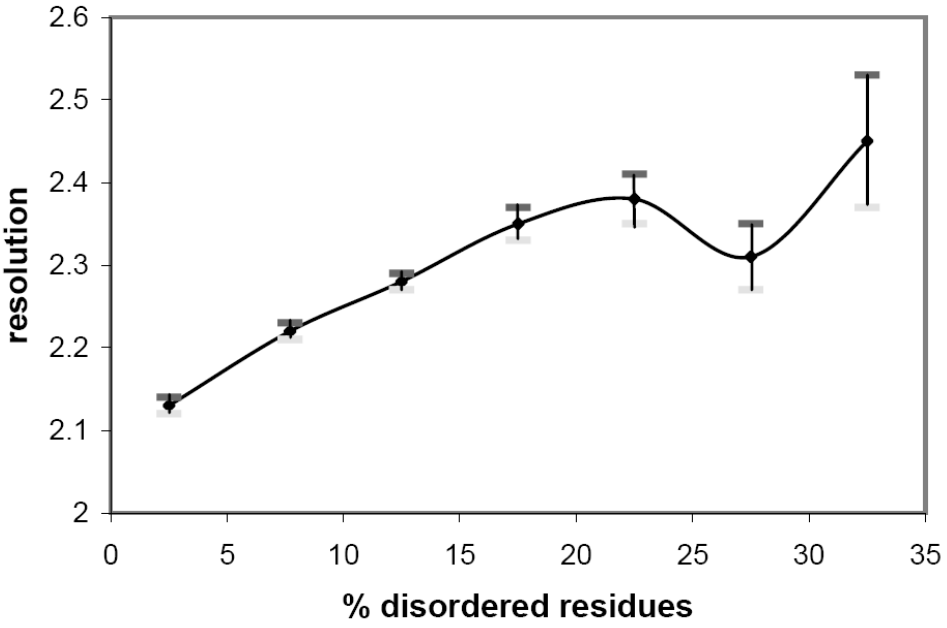


Figure 6. Dependence between the crystallographic resolution and the percentage of disordered residues observed in the crystal structures deposited in the Protein Data Bank. Vertical bars indicate the standard deviation of the mean (Kumar and Carugo 2008).

4.6. Protein quaternary status prediction

4.6.1. Introduction to protein structure

Proteins are polymers of 20 different amino acids joined by peptide bonds. Based on structure hierarchy protein structures are classified as primary, secondary, tertiary, and quaternary (Yu et al. 2006). Primary Structure refers to the linear sequence of amino acids that make up the polypeptide chain. The secondary structure of protein molecules refers to the formation of a regular pattern of twists or kinks of the polypeptide chain. The regularity is due to hydrogen bonds formed between the atoms of the amino acid backbone of the polypeptide chain. The two most common types of secondary structure are called α helix and ß pleated sheet. Tertiary structure refers to the three-dimensional globular structure formed by bending and twisting of the polypeptide chain. This process often means that the linear sequence of amino acids is folded into a compact globular structure. The folding of the polypeptide chain is stabilized by multiple weak, non-covalent interactions. Quaternary structure refers to proteins that contain more than one polypeptide chain. Each polypeptide chain in the protein is called a subunit. The subunits can contain the same polypeptide chain or different ones.

4.6.2. Quaternary structure

The concept of quaternary structure was first put forward by Bernal in 1958 (cited in Klotz et al. 1970). Many proteins self-associate into assemblies composed by two or more polypeptide chains. Protein assemblies composed of one polypeptide chain are termed as monomers and those composed of more than one polypeptide chain are called oligomers. Oligomer names are based on the number of subunits; dimers are containing two subunits, trimers containing three subunits, tetramers containing four subunits and so

on. Oligomers which have identical subunits are called homo-oligomers and those which are not are called hetero-oligomers. Association of several subunits into a protein has important consequences for its function, which is often lost if the subunits are separated. Quaternary structure complexes are involved in various biological processes, which include metabolism, signal transduction and chromosome replication.

4.6.3. Importance of quaternary structure prediction in structural biology

According to Jones and Thornton (1996), the quaternary status of proteins can be described with different types of assemblies. Permanent complexes include those proteins that only function in the complex state, and are thus obligatory, e.g. oligomeric proteins. Non-obligate complexes are built from units that exist both as part of the complex and separately in the cell e.g. enzymes and their inhibitors. Prediction of quaternary status of protein is important in structural biology before starting an experimental analysis. The structure of an isolated protein chain of a permanent and obligate hetero-oligomeric protein cannot be determined experimentally since it has low solubility and conformationally inhomogeneity. Moreover, this would anyway result into experimental artifacts. Therefore it is important to predict permanent and obligate hetero-oligomeric assembly in order to avoid selection of 'impossible targets'. Moreover quaternary structure determination through experiments is slow and expensive. However, computational methods can play a vital role in extracting valuable information rapidly that helps in structural genomics project.

4.6.4. Computational tools for quaternary structure prediction

It is generally accepted that the amino acid sequence of proteins contains all the information needed to fold the protein into three-dimensional structure (Anfinsen 1973). The association of tertiary structure of subunits depends upon the complementary 'patches' on their surfaces. The patches are buried in the interfaces formed by the subunits, thus, play a role in both tertiary and quaternary structure. This suggests that primary sequences contain quaternary structure information (Garian 2001).

Quite a few computational tools were developed to predict protein quaternary structure from amino acid sequence. Computational methods like machine learning methods were employed to predict the number of subunits from amino acid sequences, and in discrimination between quaternary statuses of protein. Garian (2001) employed the decision-tree method with a feature extraction function to discriminate between homo-dimer and non homo-dimer. Chou and Cai (2003) developed a method for identifying number of subunits in homo-oligomeric quaternary structure by implementing the pseudo amino acid composition method. Zhang et al.(2003) developed a method for discriminating between homo-dimer and non homo-dimer, using the support vector machine and the covariant discriminant algorithm taking in account the amino acid composition and auto correlation functions. Song and Tang (2004) implemented the function of degree of disagreement (FDOD) to discriminate between homo-dimers and other homo-oligomeric proteins from their primary sequence. Carugo (2007c) developed a computational method for discriminating between hetero-oligomer and non hetero-oligomer by implementing the Mahalanobis distance taking in account the amino acid

composition. With the exception of the last study, none of these studies dealt with hetero-oligomers, the target of the present study.

4.6.5. Quaternary structure and the Protein Data Bank

Jones and Thornton (1996) studied the distribution of multimeric states of proteins in the July 1993 release of the PDB. They noted about 66% was monomeric and concluded that PDB over-represents small-monomers owing to the difficulties involved in protein crystallization of supramolecular assemblies. 15% of the PDB entries were dimeric and 12% were tetrameric and the remaining adopted other oligmeric statuses.

4.7. Prediction of metalloproteins

4.7.1. Metalloproteins

The proteome of every organism requires a significant share of metal ions or metal containing cofactors to carry out its physiological function. Metalloproteins are proteins capable of binding one or more metal ions or metal containing cofactors, which are required for biological function or for the regulation of their activities or for structural purposes (Passerini et al. 2007). In *in vitro* condition, metal ions are observed to interact with unfolded polypeptide and may create local structure that initiates and directs the polypeptide folding process (Wilson et al. 2004).

Metal-binding capabilities are encoded in the amino acidic sequences and these primary sequences are related to the protein three-dimensional structure. Through genomic projects various organism genomic sequences have been annotated somehow along with metalloproteins contained in them (Andreini et al. 2004). Identification of metal binding through experimental methods is difficult and expensive. The use of bioinformatics has been extensively used to predict metal binding from amino acid sequences. Predictions of metal binding proteins are useful in structural genomics, to select proper growth medium for over-expression studies and for the easy interpretation of electron density maps.

However, the available prediction methods are either based on the knowledge of the apo-protein structure or they are restricted to few specific cases, like the metal binding of histidines/cysteines.

4.7.2. Metal-binding predictors

(i) CHED

CHED (Babor et al. 2008) predicts transition metal-binding sites in apo- proteins on the basis of their three-dimensional structure. The algorithm first uses geometric search, for a triad amino acids composed of four residues types (Cys, His, Glu, Asp) taking account of structural rearrangements upon mental binding. Machine learning algorithms (decision trees and support vector machines) were used to filter out false positive. The web server can be accessed at http://ligin.weizmann.ac.il/~lpgerzon/mbs4/mbs.cgi.

(ii) SeqCHED

SeqCHED (Levy et al. 2009) is a web-tool for predicting the metal binding sites of proteins from translated gene sequences based on remote homology templates with sequence identity between 18-100%. A metal binding prediction algorithm (based on the CHED procedure) is then applied to the three-dimensional model to identify any putative binding sites and their ligating CHED (Cys, His, Glu, Asp) residues. The web server can be accessed through http://ligin-temp.weizmann.ac.il/~ronenle/Web/SeqCHED/.

(iii) Metal detector

Metal detector server (Lippi et al. 2008) is used to predict the metal binding capacities of cysteines and hisitidines and disulfide bridges for transition metals from protein sequence. A decision tree integrates predictions from two previously developed (Disulfind and metal ligand predictor) methods. The server can be accessed through http://metaldetector.dsi.unifi.it/.

(iv) Metal ligand predictor

The metal ligand predictor (Passerini et al. 2006) identifies cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. It is based on a two-stage machine-learning approach. The first stage consists of a support vector machine trained to locally classify the binding state of single histidines and cysteines. The second stage consists of a neural network trained to refine local predictions by taking into account dependencies among residues within the same protein. The method predicts histidines as being in either of two states (free or metal bound) and cysteines in either of three states (free, metal bound, or in disulfide bridges). The method uses only sequence information by utilizing position-specific evolutionary profiles as well as more global descriptors such as protein length and amino acid composition. The dataset is available at http://www.dsi.unifi.it/passe/datasets/mbs06/dataset.tgz.

(v) Met site

The met site server (Sodhi et al. 2004) locates metal-binding regions in protein structures using a set of artificial neural network classifiers. The server uses secondary structure, solvent accessibility and distance matrices to improve the classification performance. The web server can be accessed through http://bioinf.cs.ucl.ac.uk/MetSite/MetSite.html.

(vi) GRID

GRID (Goodford 1985) is a computational procedure for determining energetically favorable binding sites on proteins on known structure. The website can be accessed at http://www.moldiscovery.com/soft_grid.php.

4.8. Structure and function of filamin

4.8.1. Introduction

The cytoskeleton plays a fundamental role in spatial organization of cells and their movement. The actin cytoskeleton of eukaryotic cells is important for the maintenance of cell shape, cell division, adhesion, motility, signal transduction, phagocytosis and protein sorting. The actin cytoskeleton is regulated by many proteins that perform different functions like actin polymerisation and cross-linking of actin filaments. In non-muscle cells, the actin cytoskeleton consists of globular monomeric actin (G-actin), which can reversibly polymerize into filamentous actin (F-actin).

The cross linking and localization of filamentous actin (F-actin) is done by several proteins, including spectrin, fimbrin, α-actinin and filamin. Proteins that cross-link F-actin are important for the maintenance of the viscoelastic properties of the cytoplasm and the generation of cell locomotion (Rivero et al. 1996). The cross linking proteins like spectrin, fimbrin and α-actinin are thought to form primarily parallel actin bundles, whereas filamins can crosslink actin filaments to form orthogonal networks to bundles depending on their concentration (Popowicz et al. 2006). The spectrin and α-actinin rod domain consist of α-helical regions whereas filamin rod domain are characterized by presence of several immunoglobulin-like folds, in which seven β strands are arranged in an antiparallel way.

4.8.2. Filamin

The name filamin refers to its filamentous colocalization with actin stress fibers. Filamins are large cytoplasmic homo-dimeric proteins that crosslink cortical actin into three-dimensional structures and give mechanical force to cells by binding to actin filaments and forming bundles or gel networks (van der Flier and Sonnenberg 2001). Monomeric chains of filamin comprise an actin binding domain (ABD) at the N-terminus, followed by 4-24 immunoglobulin Ig-like domains depending on the organism including a carboxy terminal dimerization domain (Stossel et al. 2001) as shown in figure 7.

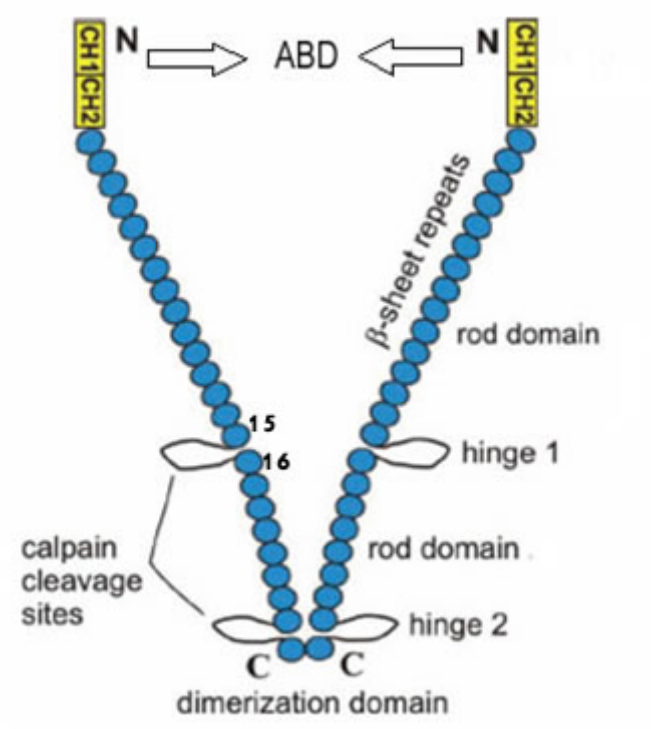


Figure 7. Overall structure of human filamin (modified from Stossel et al. 2001).

Table 5. List of interaction partners of human filamin.

Protein	*Filamin binding site*	*Function*	*Reference*
Gpibα	Ig-like domain 17-19	Platelet adhesion receptor	(Cranmer et al. 2005; Nakamura et al. 2006)
D2/D3 dopamine receptors	Ig-like domain 19	Receptor to actin anchoring	(Lin et al. 2001)
Kir2.1	Ig-like domain 23–24	Receptor to actin anchoring	(Sampson et al. 2003)
FILIP	Unknown	Downregulated by FILIP	(Nagano et al. 2004)
Furin	Unknown	Sorting, compartmentalization and stabilization	(Liu et al. 1997)
Migfilin	Ig-like domain 21	Cell adhesion structure to cytoskeleton binding	(Wu 2005; Tu et al. 2003)
RalA	Ig-like domain 24	Cytoskeleton regulation, filopodia formation	(Ohta et al. 1999)
FAP52	Ig-like domain 13–16	Unknown	(Nikki et al. 2002)
Caveolin-1	Ig-like domain 22–24	Anchoring caveolae to cytoskeleton	(Stahlhut and van Deurs 2000)
Smad	Ig-like domain 20–23	Anchoring and phosphorylation promotion	(Sasaki et al. 2001)
TRAF1, TRAF2	Ig-like domain 15–19	Anchoring and receptor internalization and recycling	(Arron et al. 2002)
CaR extracellular Ca^{2+} receptor	Ig-like domain 14–16	Receptor to actin anchoring	(Awata et al. 2001)
FOXC1	Ig-like domain 20	Nuclear scaffold	(Berry et al. 2005)
SHIP-2	Unknown	Receptor to actin anchoring	(Dyson et al. 2003)
HCN1	Ig-like domain 24	Receptor to actin anchoring	(Gravante et al. 2004)
Glutamate receptor type 7	Ig-like domain 21–22	Receptor to actin anchoring	(Enz 2002)
Calcitonin receptor	Ig-like domain 20–22	Anchoring and receptor internalization and recycling	(Seck et al. 2003)

Androgen receptor	Ig-like domain 16–24 after cleavage	Downregulates AR in nucleus	(Loy et al. 2003; Ozanne et al. 2000)
SEK-1	Ig-like domain 21–23	Tumour necrosis factor-α activation	(Marti et al. 1997)
BRCA-2	Ig-like domain 21–24 in nucleus	Promotes recovery from G2 arrest after DNA damage	(Meng et al. 2004; Yuan and Shen 2001)
Protein kinase Cα	Ig-like domain 1–3; hinge 2 to Ig-like domain 24	Scaffold for signalling pathway	(Tigges et al. 2003)
Integrin	Ig-like domain 21	Receptor to actin anchoring	(Travis et al. 2004; Nakamura et al. 2006)
Pak1	Ig-like domain 23	Ruffle formation	(Vadlamudi et al. 2002)
PEBP2/CBF	Ig-like domain 23–24	Retains PEBP2 in cytoplasm inhibiting its nuclear activity	(Yoshida et al. 2005)

Besides actin cross-linking, filamin functions involve anchoring of transmembrane proteins, membrane stabilization, interactions between cells and the extracellular matrix , scaffold for various signaling molecules and functions related to protein trafficking. Filamin proteins are reported to be interacting with more than 20 proteins as show in table 5.

4.8.3. Calponin Homology (CH) domain

The calponin homology (CH) domain is one among the many protein domains, which are shared by cytoskeletal and signaling proteins and had been identified in a number of actin binding proteins. The actin-binding domain is divided into two dissimilar CH domains (an N-terminal or type 1 CH domain and C-terminal or type 2 CH domain). It has been shown to mediate the actin-binding properties of multiple proteins. These two dissimilar CH domains are observed to show differences in binding affinities with F- actin.

Type 1 CH domains have the intrinsic ability to interact with F-actin, while type 2 CH domains do not, but contribute substantially to the interaction of the complete actin binding domain, by acting as a locator or low affinity docking site on the actin filament.

The two tandem calponin homology domains (CH1 and CH2) consist of four main α-helices of 11-18 residues, connected by two or three less regular helices (Gimona and Mital 1998). An example of CH domain of human β-spectrin is shown in figure 8.

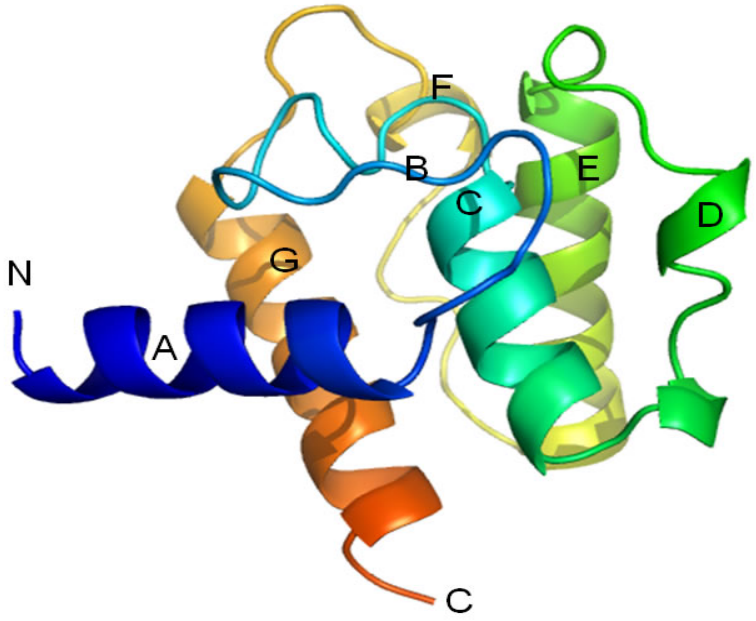


Figure 8. CH domain of human β-spectrin (PDB ID:1aa2). The architecture of the domain is dominated by four α-helices (A,C,E and G) connected by long loops and three short less regular α-helices (B,D,F) (Djinovic Carugo et al. 1997).

4.8.4. Structure of actin-binding domain

Generally, the cross-linking of actin requires at least two actin-binding sites, one per each filament. A classification of actin binding and cross-linking proteins is based on domain composition (Djinovic-Carugo et al. 2002). Proteins involved in F-actin cross-linking

generally fall into three subclasses (Gorlin et al. 1990). The first subclass has the simplest organization with two tandem ABDs on the same polypeptide chain (fimbrin, plastin); the second subclass (α-actinin, spectrin, dystrophin, plakin families, utrophin) form non-covalent dimers via a coiled-coil called a spectrin repeat (Popowicz et al. 2006); the third subclass, which includes filamins, is characterized by a dimerisation of an antiparallel seven-stranded β- barrel adopting an Ig-like fold.

The typical ABD has 250 residues and shares 20-60% of sequence identity with other ABDs in the family (e.g. filamin, spectrin, fimbrin, nesprin, and plectin). In ABD, three N-terminal α-helices form a triple helical bundle, the amino-terminal α-helix of which packs in a perpendicular orientation while a fourth carboxy-terminal long α-helix connects the two domains (Popowicz et al. 2006). From mutation studies, it has been identified that ABD family members have three potential actin-binding sites (ABS1, 2 and 3). The first α-helix of CH2 domain (ABS3) contributes to F-actin binding. The last α-helix of CH1 domain (ABS2) which has conserved hydrophobic region is crucial for binding. The amino acids of actin that participate in binding are located between residues 112-125 and 360-372.

4.8.5. Actin-binding domain of filamin

The actin-binding domain of filamin composed of two calponin homology (CH) domains: the amino- and carboxy-terminal CH1 and CH2 domains. Each CH domain consists of four main α-helices connected by long loops, and two or three shorter, less regular α-helices. Three dominant alpha helices form a triple helical bundle, against which the

amino-terminal α-helix packs in a perpendicular orientation (van der Flier and Sonnenberg 2001). An example of actin-binding domain of filamin A is shown in figure 9.

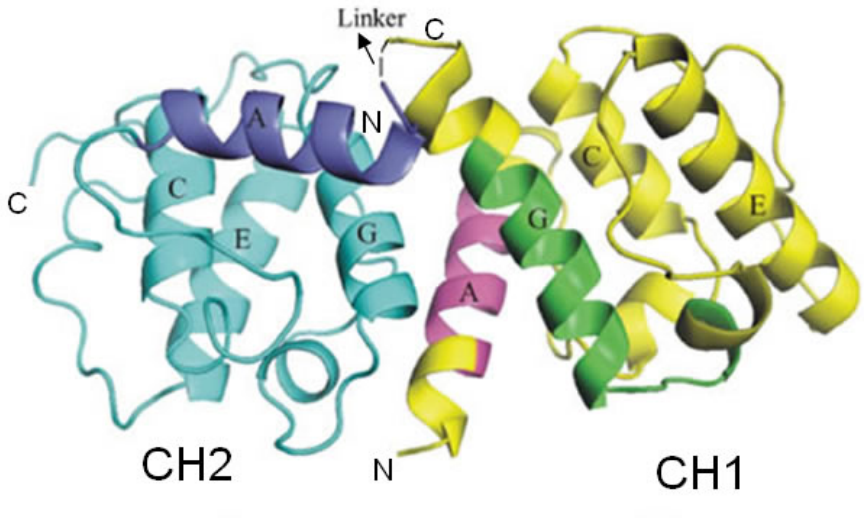


Figure 9. A cartoon representation of filamin A actin-binding domain chain A. CH1 and CH2 domain connected by the linker region. ABS1 is shown in magenta, ABS2 in green and ABS3 in blue. The architecture of the domain is dominated by four α-helices (A, C, E and G) connected by long loops (modified from Ruskamo and Ylänne, 2009).

The recent high-resolution X-ray crystal structures of the human filamin B wild type ABD with other mutant's shows that they have compact monomeric conformation for the ABD with CH1 and CH2 domains in close contact. The filamin B ABD is 242 amino acids long comprising two calponin homology (CH) domains, designated CH1 and CH2, each approximately 100 amino acids (Sawyer et al. 2009).

4.8.6. Rod region

The function of an actin-binding protein is dictated by the mechanochemical properties of its rod region. The actin-binding protein fimbrin is a monomer with multiple tandem repeats of the ABD. In fimbrin, the absence of rod region results in tight actin bundles (Goldsmith et al. 1997). Parallel and less dense formations of actin are induced by α-actinin, antiparallel homo-dimer containing four rod domains versus none in fimbrin (Puius et al. 1998; Zaman and Kaazempur-Mofrad 2004). Even more diverse are the filamins, containing a longer rod region in addition to the ABD in each subunit of the antiparallel homo-dimer.

4.8.7. Rod region of filamin

In filamin, the long rod region facilitates binding and stabilizing of actin into an orthogonal network of filaments. The human filamin rod domains consist of a β sandwich, which resembles the subtype C1 fold of the immunoglobulin family (Fucini et al. 1999). Human filamin consists of 24 rod domains composed of immunoglobulin Ig-like fold whereas *Dictyostelium discoideum* filamin (ddFLN) has six tandem repeats. The filamin repeats are interrupted by one or two flexible non-modular hinge regions one between 15 and 16 and other between 23 and 24 of Ig-like domains as shown in figure 7.

4.8.8. Dimerisation of filamin

Dimerisation is crucial for the actin cross-linking function of filamins and occurs through the most C-terminal domain. Based on the significant similarities in protein sequences it

was proposed that the dimerisation mode found in human filamin is common for all vertebrate filamins. It has been shown that the Ig-like domain 24 alone is sufficient for filamin dimerisation. The dimerization of filamin allows the formation of a V-shaped flexible structure. The crystal structure of human Filamin C Ig-like domain 23 along with small angle scattering (SAXS) study of Ig-like domain 23 and 24, shows that there is no significant involvement of Ig-like domain 23 in dimer formation (Sjekloca et al. 2007).

4.8.9. Filamin isoforms

Filamin in humans consists of three genes, (FLNA, FLNB, FLNC), which encode filamin A, B, and C, respectively (figure 10). FLNA and FLNB are expressed ubiquitously and FLNC is expressed in skeletal and cardiac muscles. Three filamin gene paralogues have been mapped on different chromosomes, FLNA is located on chromosome Xq28 (Maestrini et al. 1993); FLNB on chromosome 3pl4.3 (Chakarova et al. 2000); and FLNC on chromosome 7 (Popowicz et al. 2006). The exon-intron structure of all three human filamin gene paralogues is highly conserved, but the gene organization does not correlate with the domain structure of the proteins (Xie et al. 1998).

In human filamin, each isoform has a relative molecular mass of 280 kDa, consists of an amino-terminal actin binding domain (ABD) and 24 Ig-like domains. The three filamin proteins (filamin A, filamin B and filamin C) show 60-80% sequence identity along the entire molecule, and have divergence at the two hinge regions of the rod region. All human filamins has two unique long hinges positioned between repeats 15-16 (27 residues) and 23-24 (35 residues) that are postulated to be flexible (van der Flier and

Sonnenberg 2001). However, only filamin C contains an 81 amino acid insertion in repeat 20, not present in filamin A or filamin B (Krakow et al. 2004).

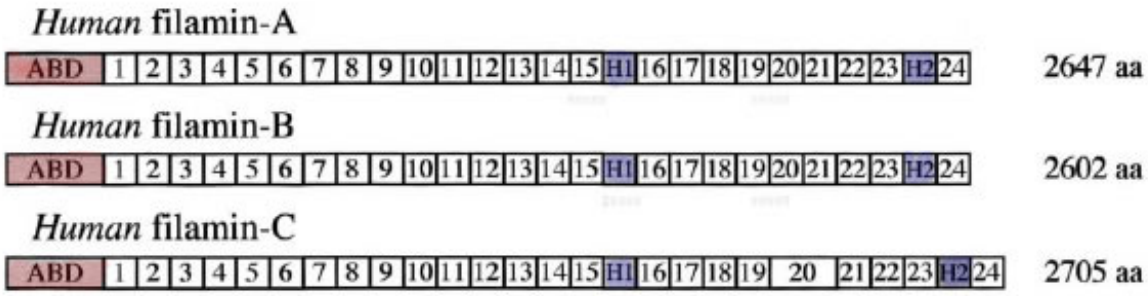


Figure 10. The architecture of human filamin isoforms. Indicated are the actin-binding domains (ABD) at the N-terminus; the Ig-like domains numbered from 1 to 24; the hinge-1 (H1) and the hinge 2 (H2) regions. (modified from Stossel et al. 2001).

4.8.10. Filamin functions

Filamins have diverse functions: (i) organising the actin cytoskeleton; (ii) providing a link between extra cellular matrix, plasma membrane and actin cytoskeleton through interaction with a number of transmembrane receptors; (iii) serving as a platform for a variety of signaling molecules and thus playing an important role in signal transduction between the cell membrane and cell interior (Robertson 2005) (figure 11). Moreover, filamin mutations have been related to several human diseases affecting the brain, bone and cardiovascular system, as well as muscle fibers (Stossel et al. 2001; Tseng et al. 2004).

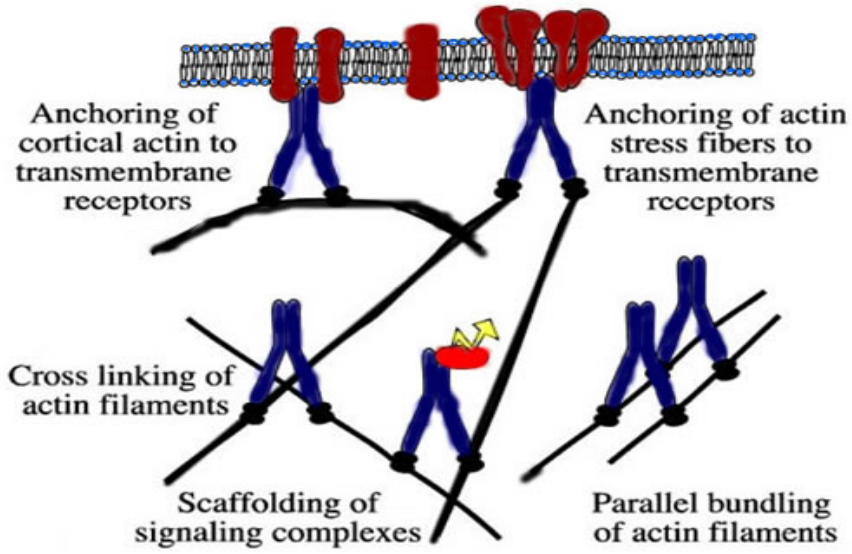


Figure 11. Filamin biological functions (modified from van der Flier and Sonnenberg 2001).

(i) Role of filamin and its interaction with actin and actin organisation

Filamin induces the formation of gelatinous actin, by cross-linking actin filaments into orthogonal networks. It is evident that filamin must dimerise but the mechanism of how filamin promotes actin filament branching is not completely understood (Calderwood et al. 2001).

The organization of actin inside the cell depends on the intracellular ratio of filamin to actin and the formation of bundles or networks is dependent on the structure of filamin variants. Through *in vitro* experiment with recombinant filamin lacking hinge region 1, it is demonstrated that filamin forms linear stiffening and a consequent of breakage of the cross-linked actin structures at much lower stress (Gardel et al. 2006).

(ii) Role of filamin and its interaction with transmembrane proteins

Filamins provide stabilization to cell membrane, maintain cell-cell and cell-matrix connections by association with transmembrane proteins such as β-integrins (Sampson et al. 2003), glycoprotein (GP)Ib-IX-V transmembrane complex (He et al. 2003), ion channels like Kir2.1 (Ohta et al. 1999), insulin receptor (Stossel et al. 2001), and small GTPases and related proteins like RalA, RhoA,Rac1 (Vadlamudi et al. 2002). Filamin plays an important role in transmembrane signaling, regulating cell adhesion and cell shape regulation through the association of integral membrane receptors with cytoskeleton.

(iii) Role of filamin in signal transduction

Filamin serves as a scaffold for many intercellular signaling molecules and is involved in the regulation of signaling molecules. Although filamins are linked to a number of signaling pathways, it is unclear how the signaling reactions affect filamin function (Sells et al. 1997). Filamin through interaction with p21-activated kinase (Pak21) (Sells et al. 1999) results in the formation of ruffles, lamellipodia and filopodia (Ohta et al. 2006) and regulates cell motility (Shifrin et al. 2009) which is indirectly involved in the regulation of actin cytoskeleton.

4.8.11. Regulation of filamins

Filamin functions are regulated at many levels by binding of phospholipids (Rosenberg et al. 1981), phosphorylation by serine/threonine kinase (Chen and Stracher 1989), and by

calcium binding (Ohta and Hartwig 1995). The phosphorylation and dephosphorylation process regulate the interaction of filamin with other cytoskeletal elements.

The phoshphorylation of filamin by protein kinase A increases its resistance to calpain cleavage (Garcia et al. 2006); and phosphorylation by calcium /calmodulin-dependent protein kinase II (CaM kinase II) decreases its actin-binding affinity (Nakamura et al. 2005).

Dephosphorylation by calcineurin, a calcium/calmodulin dependent theronin/serine phosphatase, protects filamin in platelets from calpain degradation. *In vivo*, assembly and disassembly of the actin cytoskeleton in platelets is controlled by the intracellular concentration of free calcium. Recently, the first mechanism for the regulation of filamin interaction with F-actin has been proposed, providing an explanation for why a large part of cellular filamin stays free of F-actin *in vivio* (Nakamura et al. 2005). It involves direct interaction between filamin and calcium/calmodulin, which dissociates F-actin from filamin and inhibits its ability to crosslink actin filaments.

5. Methods

5.1. Conformational disorder prediction

Protein sequences were downloaded from the Disprot database (http://www.disprot.org/) release 3.3 (Sickmeier et al. 2007), which contains information about conformationally disordered proteins. The data, downloaded in August 2006, contain about 458 proteins. Each residue of these 458 proteins is labeled according to its conformational status: ordered, disordered, and unknown. 12 individual predictors were used (table 4) and each prediction method produces binary results: a residue can be predicted to be conformationally ordered or disordered. More details are available in section 8.1.

5.2. Prediction of quaternary status of protein

5.2.1. Datasets

Protein sequences were downloaded from the UniprotKB database (Wu et al. 2006) according to their quaternary status (using the keywords monomer, homo-dimer, homo-trimer, homo-tetramer, homo-pentamer, homo-hexamer, hetero-dimer, hetero-trimer, hetero-tetramer, hetero-pentamer or hetero-hexamer). The downloaded sequences contain about 11096 monomers, 43088 homo-oligomers and 13669 hetero-oligomers. Sequences containing non-standard residues were omitted, together with membrane proteins (identified by using the server http://www.cbs.dtu.dk/services/TMHMM/), and the sequence redundancy was reduced with the program cd-hit (40% maximal identity) (Li & Godzik, 2006), resulting in 1404 monomeric, 2982 homo-oligomeric and 1444 hetero-oligomeric proteins. A total of 5830 protein chains were examined. Each protein

sequence was represented by the percentage of occurrence of each of the 20 natural amino acids.

5.2.2. Parameterization of the protein dimension

Not surprisingly, the amino acid composition depends on the dimension of the proteins (Carugo 2008d). Both the number of residues exposed to the solvent (which tend to be polar) and the number of residues buried within the protein interior (which tend to be apolar) increase if the total number of the residues increases. However, the protein surface increases less than the interior, just as the surface of a sphere increases less than the sphere volume if the radius of the sphere increases.

Table 6. Number of protein chains within each of the 11 subgroups of protein dimensions.

Range of number of residues	*Monomeric protein chains*	*Homo-oligomeric protein chains*	*Hetero-oligomeric protein chains*
0-100	111	127	95
100-200	189	381	222
200-300	232	576	208
300-400	238	568	182
400-500	196	457	176
500-600	142	265	113
600-700	92	158	103
700-800	65	98	82
800-900	40	63	66
900-1000	35	55	45
>1000	64	234	152

For this reason, the data sets were divided into 11 subsets: one containing proteins with less than 100 residues, one with 100-200 residue proteins, one with 200-300 residue proteins, one with 300-400 residue proteins, one with 400-500 residue proteins, one with

500-600 proteins, one with 600-700 residue proteins, one with 700-800 residue proteins, one with 800-900 residue proteins, one with 900-1000 residue proteins, and the last one with proteins containing more than 1000 residues (table 6).

5.2.3. Predictions

We used several approaches provided by the freely available package WEKA (http://sourceforge.net/projects/weka; Witten & Frank, 2005). They include Naive Bayes, support vector machines, trees, meta and rules classifiers etc.

5.2.4. Prediction validation

All predictions were validated with a tenfold cross-validation. The reliability of the predictions was calculated with the following quantities: tp (true positives) is the number of correctly predicted hetero-oligomeric proteins, tn (true negatives) is the number of correctly predicted non-hetero-oligomeric proteins, fp (false positives) is the number of proteins that were predicted to be hetero-oligomeric although they are not, and fn (false negatives) is the number of hetero-oligomeric proteins that are predicted to be non-hetero-oligomeric although they are. On the basis of these quantities, we computed the sensitivity, the specificity, the accuracy, the Mathews correlation and the probability excess, as described in section 4.4.5.4

5.3. Prediction of metalloproteins

5.3.1. Metal/non-metal datasets

All the protein sequences were downloaded from the UniProt database (Wu et al. 2006) available at http://www.uniprot.org/. The downloaded sequences, annotated as metal-

containing, were grouped into eight subsets. Each of the subsets, containing one of the metal species viz., calcium, cobalt, copper, iron, magnesium, manganese, nickel and zinc was considered to be metal- containing while all other entries were considered to be metal-free. Redundant sequences were removed with the cd-hit program (Li and Godzik 2006) at the 50% level of percentage of identity, analogous by the UniRef 50 list available at the UniProt database.

This resulted in eight data sets containing 186 calcium-containing proteins, 69 cobalt-containing proteins, 215 copper-containing proteins, 315 iron-containing proteins, 961 magnesium-containing proteins, 386 manganese-containing proteins, 74 nickel-containing proteins and 1716 zinc-containing proteins. All proteins containing calcium, cobalt, copper, magnesium, manganese, nickel or zinc were then subtracted from the UniRef50 list, resulting in a collection of 1,640,922 non-metalloproteins.

5.3.2. Selection of variables

A Simplified amino acid alphabet of 18 characters was used (table 7).It is based on three independent aminoacid classifications, viz., (i) the conformational similarity proposed by Chakrabarti and Pal (Chakrabarti and Pal 2001) to describe the conformational similarity between the 20 amino acids based on torsion angles, which contains seven clusters: [CMQLEKRA], [P], [ND], [G], [HWFY], [S] and [TIV].

(ii) The fold-dependent conservation, proposed by Murphy et al (Murphy et al. 2000) on the basis of the BLOSUM 50 matrix, that groups together on the basis of the possibility

of foldable structures and consists of the clusters: [P], [KR], [EDNQ], [ST], [AG], [H], [CILMV] and [YWF] and

(iii) The hydrophobicity proposed by Rose et al. (Rose et al. 1985) which consists of the following groups: [CFILMVW],[AG],[PH],[EDRK] and [NQSTY].

Table 7. The 18 variables, obtained by merging three simplified alphabets of amino acid residues, used to represent protein sequences.

Variable	***Residues***
V1	CMQLEKRA
V2	P
V3	ND
V4	G
V5	HWFY
V6	S
V7	TIV
V8	CFILMVW
V9	AG
V10	PH
V11	EDRK
V12	NQSTY
V13	FWY
V14	CILMV
V15	H
V16	ST
V17	EDNQ
V18	KR

Some of the cluster [P] and [AG] which are present in more than one simplified alphabet were considered only once. These results in 18 variables and the proteins are represented with their percentage of observations.

5.3.3. Random forest predictions

We used several machine learning methods available within the package WEKA (http://sourceforge.net/projects/weka/; Witten & Frank, 2005). The best results were obtained by using 'random forest trees'. A random forest is an ensemble of decision trees, built by random selection of subsets of the data and by using random subsets of the variables. 30 random trees were built in each prediction run and the final prediction was determined by the majority rule.

5.3.4. Prediction validation

All predictions were validated with a tenfold cross-validation. When the predictor was focused on the problem of distinguishing proteins containing a certain type of metal ion from proteins that do not contain any type of metal, three runs were performed, each time using non-superposing and balanced data sets. It is important that both sets contain the same number of proteins; otherwise, several figures of merit that are commonly used to monitor the prediction reliability would be seriously biased.

The reliability of the predictions was monitored with the following quantities. If a protein of type 1 must be distinguished from a protein of type 2, a prediction was considered to be a true positive (tp) if type 1 was correctly predicted; it was considered to be a true negative (tn) if type 2 was correctly predicted; it was considered to be a false negative (fn) if a type 1 protein was predicted to be a type 2 protein; and it was considered to be a false positive (fp) if a type 2 protein was predicted to be a type 1 protein. Consequently,

the following figures of merit, the sensitivity, the specificity, the accuracy, the Mathews correlation and probability excess are computed as described in section 4.4.5.4

5.4. Filamin bioinformatic characterization

5.4.1. Dataset

5.4.1.1. Filamin datasets of individual domains

Human filamin (filamin A, filamin B and filamin C) sequences were downloaded from the UniProt database available at http://www.uniprot.org/. For each filamin, 26 sets were constructed for the CH1, CH2 and the Ig-like domains 1-24. Domain boundaries were taken directly from the UniProt annotations.

5.4.1.2. Filamin datasets of large segments

For each filamin (filamin A, filamin B and filamin C), 26 sets were constructed containing increasing segments by combining each individual domain like CH1-CH2, CH1-CH2-domain 1, and CH1-CH2-Ig-like domain 1- Ig-like domain 2 till CH1-CH2-Ig-like domain 1... Ig-like domain 24.

5.4.1.3. Filamin test set for quaternary status prediction

For quaternary status prediction, individual domains containing 26 sets were represented by the percentage of occurrence of each of the 20 natural amino acids.

5.4.1.4. Filamin test set for metal binding prediction

For metal binding prediction, individual domains and large segments were represented with simplified amino acid alphabet as described in section 5.3.2 (see table 7).

5.4.2. Homology models of all the domains of filamin

For conformational disorder prediction of filamin protein, homology models of all domains of filamin were generated using the Modeller software version 9v5 (Fiser and Sali 2003). Templates were chosen using similarity search which is available in the PDB (http://www.pdb.org/pdb/search/advSearch.do?st=SequenceQuery). The structure which has high sequence identity, less gap between target and the template sequence was selected (see table 8).

Table 8. Homology models of human filamin isoforms. The table shows the template taken from PDB and the crystallographic resolution and sequence identity (%) between query and template. The resolution is not indicated for the NMR structures.

Domain	*Template*			*PDB ID*			*Resolution (Å)*			*Sequence identity (%)*		
	A	B	C	A	B	C	A	B	C	A	B	C
CH1	CH	CH	CH	2eyi	2eyi	2eyi	1.7	1.7	1.7	43.9	43.9	43.9
CH2	CH	CH	CH	2eyi	2eyi	1wku	1.7	1.7	1.6	26.7	24.7	27.7
Ig-like domain 1	F14_C	F14_C	F14_C	2d7m	2d7m	2d7m	NA	NA	NA	31.3	32.3	36.3
Ig-like domain 2	F14_C	F14_C	F14_C	2d7m	2d7m	2d7m	NA	NA	NA	38.3	38.7	37.3
Ig-like domain 3	F12_B	F12_B	F15_B	2dic	2dic	2dmb	NA	NA	NA	38.5	36.0	37.1
Ig-like domain 4	F9_B	F9_B	F9_B	2di9	2di9	2di9	NA	NA	NA	51.6	45.1	50.5
Ig-like domain 5	F14_C	F14_C	F14_C	2d7m	2d7m	2d7m	NA	NA	NA	41.2	38.1	41.2
Ig-like domain 6	F13_B	F13_C	F14_C	2dj4	2dj4	2d7m	NA	NA	NA	26.2	27.1	33.0
Ig-like domain 7	F14_B	F13_B	F13_B	2e9j	2dj4	2dj4	NA	NA	NA	41.4	43.4	39.3
Ig-like domain 8	F11_B	F22_B	F12_B	2dib	2eeb	2dic	NA	NA	NA	27.0	27.0	28.1
Ig-like domain 9	F9_B	F9_B	F9_B	2di9	2di9	2di9	NA	NA	NA	75.2	100	82.7
Ig-like domain 10	F10_B	F10_B	F10_B	2dia	2dia	2dia	NA	NA	NA	51.5	96.8	63.1
Ig-like domain 11	F11_B	F11_B	F11_B	2dib	2dib	2dib	NA	NA	NA	65.0	100	60.0
Ig-like domain 12	F12_B	F12_B	F12_B	2dic	2dic	2dic	NA	NA	NA	65.6	97.8	69.8
Ig-like domain 13	F13_B	F13_B	F13_B	2dj4	2dj4	2dj4	NA	NA	NA	68.0	96.8	71.8
Ig-like domain 14	F14_C	F14_B	F14_C	2d7m	2e9j	2d7m	NA	NA	NA	76.2	100	95.8
Ig-like domain 15	F15_B	F15_B	F15_B	2dmb	2dmb	2dmb	NA	NA	NA	66.6	100	59.5
Ig-like domain 16	F16_C	F16_B	F16_C	2d7n	2ee9	2d7o	NA	NA	NA	64.3	91.9	81.1
Ig-like domain 17	F17_A	F17_B	F17_C	2aav	2eea	2dmc	NA	NA	NA	97.7	100	97.8
Ig-like domain 18	F18_B	F18_B	F18_B	2dmc	2dmc	2dmc	NA	NA	NA	61.7	100	64.3
Ig-like domain 19	F19_B	F19_B	F19_B	2di8	2di8	2di8	NA	NA	NA	71.1	98.9	68.8
Ig-like domain 20	F20_B	F20_B	F20_B	2dlg	2e9i	2dlg	NA	NA	NA	60.6	97.8	69.8
Ig-like domain 21	F21_B	F21_B	F21_B	2ee6	2ee6	2ee6	NA	NA	NA	82.7	98.9	84.9
Ig-like domain 22	F22_B	F22_B	F22_C	2eeb	2eeb	2d7p	NA	NA	NA	79.7	97.8	98.9
Ig-like domain 23	F23_B	F23_B	F23_C	2eec	2eec	2nqc	NA	NA	NA	75.2	100	97.8
Ig-like domain 24	F24_B	F24_B	F24_C	2eed	2eed	1vo5	NA	NA	1.4	68.4	97.8	96.8

6. Results and Discussion

6.1. Prediction of conformational disorder

Here we summarize a work that was published (Kumar and Carugo 2008). The copy of the publication is attached (section 8.1). It describes a consensus approach based on various prediction methods, the performance of which is significantly better than that of each individual predictor. The only necessary input is the amino acid sequence of the protein. Each prediction algorithm, which is freely available must be used separately (table 4) and its results (p_indi), which is +1/-1 for a residue that is predicted to be disordered/ordered, must be inserted into equation (1), together with the coefficients (X_i) reported in table 9 which were optimized by least-square minimization.

Table 9. Optimal values of the coefficients X_i to be used to compute the p_cons values.

Method	***X***
DISEMBL_hot_loops	-0.101
DISEMBL_loops	0.377
DISEMBL_remark465	-0.172
DISOPRED	0.048
DRIPRED	0.096
FOLDINDEX	0.262
GLOBPLOT_B	-0.199
GLOBPLOR_r	0.162
IUPRED_L	0.041
IUPRED_S	-0.126
PRELINK	0.078
RONN	0.141

If the value of p_cons is positive, the residue is predicted to be disordered and if it smaller than zero, the residue is predicted to be ordered.

$$p\_cons = \sum_{i=1}^{12} X_i \,.\, p\_indi_i \quad \text{---- (1)}$$

This can easily be done for each residue and, as a consequence, it is possible to reach a global picture of the conformational status of the protein.

Table 10. Performance of the new consensus prediction method compared to the individual prediction techniques.

Method	***Sensitivity***	***Specificity***	***Accuracy***	***Probability excess***
CONSENSUS	0.833	0.968	0.814	0.801
DISEMBL_HOT_LOOPS	0.481	0.974	0.494	0.455
DISEMBL_LOOPS	0.761	0.966	0.747	0.727
DISEMBL_REMARK465	0.409	0.977	0.428	0.385
DISOPRED	0.568	0.994	0.586	0.562
DRIPRED	0.640	0.975	0.642	0.615
FOLDINDEX	0.688	0.981	0.691	0.669
GLOBPLOT_B	0.421	0.990	0.445	0.410
GLOBPLOR_R	0.589	0.979	0.597	0.568
IUPRED_L	0.609	0.993	0.624	0.602
IUPRED_S	0.529	0.996	0.550	0.524
PRELINK	0.512	0.970	0.521	0.483
RONN	0.634	0.985	0.642	0.618

This new prediction method, which is essentially a weighted consensus approach, performs quite well, better than any individual prediction algorithm, as shown in table 10. It can be seen that this consensus method of prediction is very accurate, with all the figures of merit (sensitivity-83%, specificity-96%, accuracy-81%, probability excess-80%) larger than 80%. This is impossible by using individual predictors.

6.2. Prediction of quaternary status of proteins

Amino acid compositions have been widely used in a number of computational methods. In the present work, it was used to predict the quaternary status of protein chains, which can be monomeric, homo-oligomeric or hetero-oligomeric. It is crucial to be able to predict if a protein chain is a part of a permanent and obligate hetero-oligomeric assembly, since the structure of this chain, is unsuitable for structure determination. We have shown that discrimination of hetero-oligomeric from monomeric and homo-oligomeric proteins by using machine learning methods with high reliability on the basis of amino acid sequences.

6.2.1. Selection of the best machine learning algorithm

To determine the best machine learning algorithm for prediction of protein quaternary status, we have studied several classifiers which are available in the WEKA software package. We compared them by using their default parameters with the dataset of 11 subgroups (table 6). The performance of the algorithms was measured using sensitivity, specificity, accuracy, Matthews correlation coefficients (MCC) in a 10-fold cross-validation analysis. Based on the accuracy, the best-performing algorithm was the Sequential Minimal Optimization (SMO) algorithm which is a support vector Machine (SVM). The SVM is a learning machine for two-group classifications problems that transforms the attribute space into multidimensional feature space using a kernel function to separate dataset instances by an optimal hyper plane. SMO implements the sequential minimal optimization algorithm for training a support vector classifier, using polynomial kernel.

Table 11. Prediction performance of the classifiers in discriminating hetero-oligomeric from monomeric and homo-oligomeric proteins.

Classifier	*Sensitivity*	*Specificity*	*Accuracy*	*Matthews C.C*	*Probability excess*
Function-SMO	0.926	0.768	0.741	0.147	0.694
Trees–REP tree	0.962	0.751	0.739	0.054	0.713
Trees – J48	0.883	0.782	0.739	0.154	0.664

In discriminating hetero-oligomers from monomers and homo-oligomers, SMO classifier achieved an overall predictive accuracy of 74.1% (table 11). The next two top algorithms are Trees-REP tree and Trees-J48. Decision trees classify instances by sorting them down the tree from the root node to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. All of these algorithms contain automatic attribute selection for optimal performance. The performance of the decision trees classifier such as Trees-Rep and Trees-J48 in discriminating hetero-oligomers from monomers and homo-oligomers, resulted with predictive accuracy close to 73.9% (table 11). It is wise to consider several figures of merit since they monitor different aspects of the prediction reliability. From the table 11, it is observed that the function-SMO algorithm resulted in high accuracy while the other algorithm Trees-REP tree resulted in a slightly higher probability excess and in a lower Matthews correlation coefficient. Also, Tree-J48 algorithm resulted in comparable values of Matthews correlation coefficient and in lower values of probability excess. Despite the slight variations in the figures of merit, in this study, it is shown that

discrimination of hetero-oligomeric protein from monomeric and homo-oligomeric proteins is possible with high reliability by employing machine learning methods.

6.2.2. The reliability of prediction among 11 subgroups of protein chains

The reliability of prediction among 11 subgroups of proteins was only slightly variable (figure 12). The accuracy of discrimination of hetero-oligomeric proteins from monomeric and homo-oligomeric proteins ranged from 74% (proteins with 100-200 residues) to 80% (proteins with 500-600 residues). The maximum prediction accuracy (above 80%) is between the proteins with 300-400 and 500-600 residues. The variations in performance among the 11 subgroups is not surprising given the relative dearth of Uniprot entries that have a proper annotation of the quaternary status and will probably decrease in future, when database dimensions increase and their annotations improve.

6.2.3. Performance comparison to other prediction method

We have compared our classifier with other prediction method which was published a few years back (Carugo et.al.2007c); it uses Mahalanobis distance in discrimination of hetero-oligomeric from monomeric and homo-oligomeric proteins. The performance reported in the published paper in discriminating between hetero-oligmeric and non-hetero-oligomeric proteins is 78% accuracy, Matthews correlation coefficient is 0.480 and probability excess is 0.517. In comparison, our classifier achieved slightly lesser accuracy about 74%, but higher value in probability excess (0.694). Despite the different methods used, our classifier performed like the other predictor.

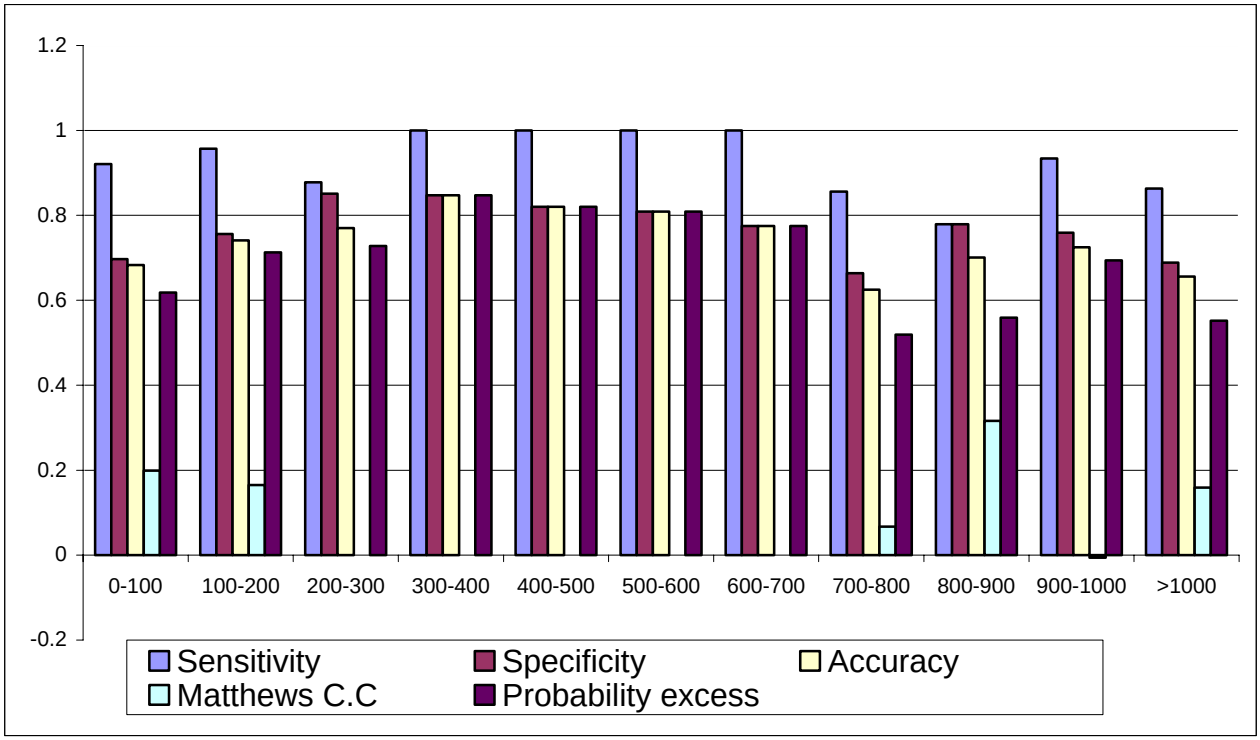


Figure 12. Prediction reliability among 11 subgroups of protein chains of different dimension (function SMO classifier).

6.3 Prediction of metalloprotein

The use of a series of simplified amino acid alphabets (table 7) allows one to identify which proteins require metal ions and which type of metal is up taken. It can be seen that metalloproteins can be identified, though the accuracy of these predictions is rather variable ranging from 69% for zinc to 90% for nickel (table 12).

Moreover, prediction performance was studied by feature selection method by removing one variable at a time and maintaining the highest value in performance indices. Measurements are removed until there is an unacceptable degradation in system performance (Guyon & Gunn, 2006). Feature selection eliminates noisy and redundant

features, the learning process is then accelerated and the accuracy of learning algorithms may be improved.

To select an optimal subset of variables, we first analyzed how individual attributes from the initial set of 18 variables, contributed to predictive accuracy. For feature selection, we employed the wrapper approach as it uses the learning algorithm to test all existing feature subsets. Wrapper methods search through the space of feature subsets and calculate the estimated accuracy of a single learning algorithm for each feature that can be added to or removed from the feature subset.

We used a backward strategy (by starting with the full set and deleting attributes one at a time) for searching the feature space. For e.g. cobalt metal binding protein can be discriminated from non-metal ions with all 18 variables with the accuracy of 85% (see figure 13). In can be seen that, on removing variable v14 (CILMV) from the subset, the accuracy of the predictor improves from 85% to 87%. After removing of variables v8 (CFILMVW), v3 (ND), v17 (EDNQ), v10 (PH), v16 (ST) the accuracy values are in the range from 86% to 87%. There is drastic decrease in accuracy of the classifier by removing the variable v12 (NQSTY) to 84%. No further reduction of the set was possible, as the performance of random forest classifier dropped if any further attributes was eliminated.

It can be seen that accuracy of prediction of metal binding proteins can be improved (for e.g. calcium from 74% to 77%, cobalt from 83% to 85%, nickel from 69% to 77%) by

elimination of certain noisy features, up to certain limit and further improvement is then impossible (table 12).

Table 12. Performance of metalloprotein prediction against proteins that lack metal ions. Separate tables are provided for different types of metal cations.

Calcium

Variable removed	***Avg.sensi***	***Avg.speci***	***Avg.preci***	***Avg.accur***	***Avg.mcc***	***Avg.pexcess***
None	0.769	0.738	0.746	0.754	0.507	0.507
P	0.783	0.758	0.763	0.770	0.541	0.541
EDNQ	0.788	0.751	0.760	0.770	0.541	0.541
EDRK	0.796	0.758	0.767	0.777	0.554	0.553
PH	0.785	0.756	0.762	0.770	0.541	0.541
CILMV	0.801	0.754	0.765	0.777	0.556	0.555
AG	0.790	0.749	0.759	0.770	0.539	0.539
CFILMVW	0.789	0.765	0.771	0.777	0.554	0.553
NQSTY	0.785	0.767	0.771	0.776	0.552	0.551
CMQLEKRA	0.780	0.765	0.769	0.772	0.545	0.544

Cobalt

Variable removed	***Avg.sensi***	***Avg.speci***	***Avg.preci***	***Avg.accur***	***Avg.mcc***	***Avg.pexcess***
None	0.884	0.823	0.832	0.853	0.708	0.707
CILMV	0.903	0.842	0.851	0.872	0.747	0.745
CFILMVW	0.899	0.837	0.847	0.868	0.737	0.736
ND	0.894	0.828	0.838	0.861	0.724	0.722
EDNQ	0.884	0.833	0.839	0.858	0.717	0.717
PH	0.894	0.847	0.853	0.87	0.741	0.740
ST	0.903	0.837	0.846	0.87	0.742	0.741
NQSTY	0.860	0.833	0.837	0.846	0.693	0.692

Copper

Variable removed	***Avg.sensi***	***Avg.speci***	***Avg.preci***	***Avg.accur***	***Avg.mcc***	***Avg.pexcess***
None	0.746	0.815	0.802	0.781	0.563	0.561
AG	0.762	0.809	0.799	0.786	0.571	0.571
CMQLEKRA	0.794	0.804	0.802	0.799	0.599	0.599
NQSTY	0.779	0.814	0.808	0.796	0.593	0.592
EDNQ	0.796	0.797	0.796	0.796	0.592	0.592
CFILMVW	0.785	0.803	0.799	0.794	0.588	0.588
TIV	0.785	0.798	0.795	0.792	0.583	0.583
PH	0.774	0.801	0.796	0.788	0.576	0.576

Iron

Variable removed	***Avg.sensi***	***Avg.speci***	***Avg.preci***	***Avg.accur***	***Avg.mcc***	***Avg.pexcess***
None	0.772	0.74	0.747	0.756	0.512	0.512
NQSTY	0.778	0.731	0.742	0.754	0.509	0.509
S	0.786	0.727	0.741	0.757	0.514	0.513
PH	0.786	0.724	0.739	0.755	0.511	0.510
CMQLEKRA	0.785	0.72	0.736	0.753	0.507	0.506
CFILMVW	0.787	0.734	0.749	0.761	0.523	0.522
AG	0.790	0.720	0.737	0.755	0.511	0.510
TIV	0.780	0.725	0.740	0.753	0.507	0.506
HWFY	0.790	0.735	0.748	0.762	0.525	0.525

Magnesium

Variable removed	***Avg.sensi***	***Avg.speci***	***Avg.preci***	***Avg.accur***	***Avg.mcc***	***Avg.pexcess***
None	0.766	0.714	0.725	0.740	0.481	0.480
ST	0.779	0.714	0.731	0.746	0.494	0.493
ND	0.774	0.720	0.734	0.747	0.494	0.493
NQSTY	0.767	0.717	0.730	0.742	0.485	0.484
S	0.772	0.711	0.727	0.742	0.484	0.483
HWFY	0.770	0.716	0.730	0.743	0.487	0.486
PH	0.777	0.709	0.727	0.743	0.487	0.486
CMQLEKRA	0.775	0.708	0.726	0.741	0.484	0.483

Manganese

Variable removed	*Avg.sensi*	*Avg.speci*	*Avg.preci*	*Avg.accur*	*Avg.mcc*	*Avg.pexcess*
None	0.729	0.647	0.674	0.688	0.378	0.377
FWY	0.731	0.717	0.746	0.734	0.474	0.474
EDNQ	0.741	0.656	0.682	0.698	0.398	0.396
CMQLEKRA	0.750	0.647	0.679	0.698	0.399	0.397
AG	0.750	0.643	0.677	0.697	0.396	0.394
S	0.739	0.660	0.684	0.700	0.400	0.399

Nickel

Variable removed	**Avg.sensi**	**Avg.speci**	**Avg.preci**	**Avg.accur**	**Avg.mcc**	**Avg.pexcess**
None	0.945	0.869	0.877	0.907	0.817	0.814
EDRK	0.950	0.887	0.893	0.918	0.838	0.837
G	0.931	0.892	0.895	0.917	0.824	0.823
NQSTY	0.923	0.887	0.890	0.905	0.810	0.810
ST	0.941	0.878	0.884	0.909	0.821	0.819
EDNQ	0.936	0.865	0.872	0.900	0.803	0.801
FWY	0.918	0.860	0.867	0.889	0.780	0.778
HWFY	0.931	0.865	0.872	0.898	0.800	0.800
TIV	0.927	0.869	0.875	0.898	0.797	0.796

Zinc

Variable removed	*Avg.sensi*	*Avg.speci*	*Avg.preci*	*Avg.accur*	*Avg.mcc*	*Avg.pexcess*
None	0.740	0.640	0.672	0.690	0.382	0.380
HWFY	0.751	0.638	0.675	0.695	0.391	0.389
CMQLEKRA	0.750	0.636	0.673	0.692	0.386	0.384
AG	0.747	0.638	0.673	0.693	0.388	0.385
ST	0.743	0.644	0.676	0.693	0.389	0.387
EDNQ	0.743	0.636	0.671	0.689	0.381	0.379

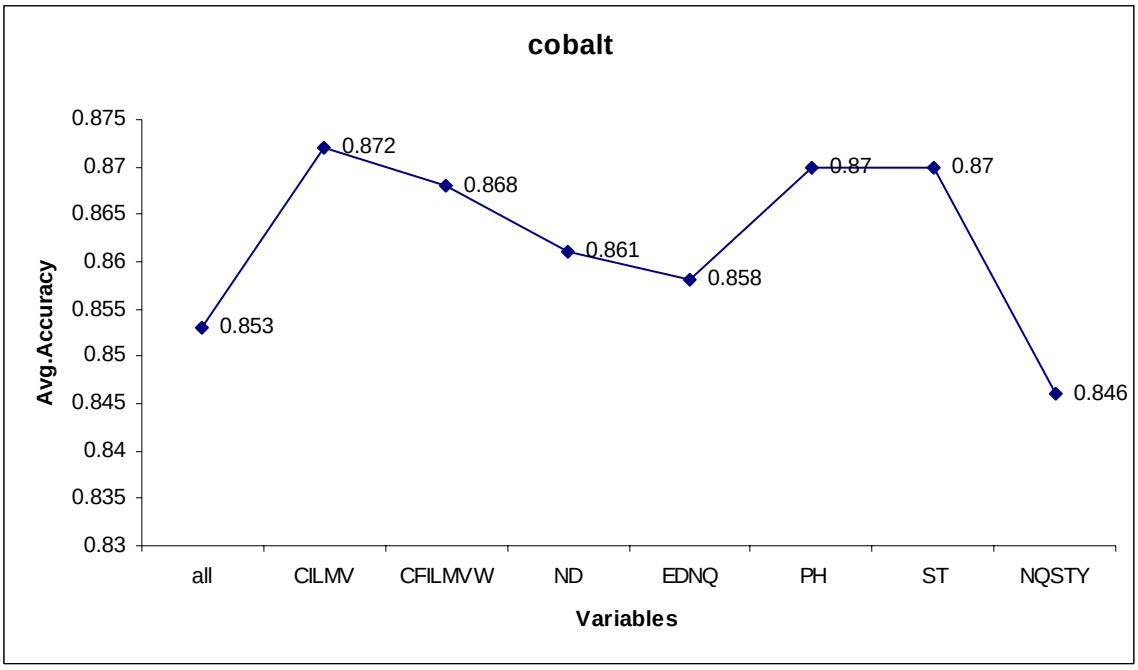


Figure 13. The performance graph of the Random forest classifier using feature selection (10-fold cross-validation).

According to this backward strategy of feature selection it can be observed that the prediction performance can be slightly improved. Some common variables rejected are v14 (CILMV) in calcium and cobalt, v8 (CFILMVW) in copper and iron.

6.3 Filamin bioinformatic characteriztion

6.3.1. Conformational disorder prediction

It is commonly assumed that a protein must attain a stable, folded conformation in order to carry out its specific biological function. However, it was recently shown that several proteins do not assume a well defined and stable three-dimensional structure but are natively unfolded. The techniques for predicting conformational disorder are extremely

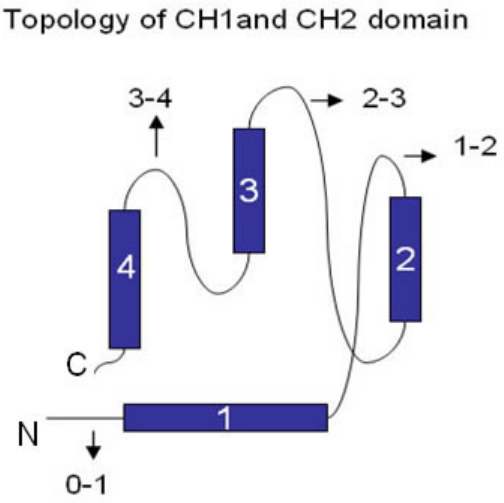
important in structural biology, where they are becoming routine filters in the pipeline of finding suitable targets to be analyzed.

The prediction of conformational disorder for individual domains of filamin was done with the consensus method (Kumar and Carugo 2008) summarized in section 6.1. Moreover the predicted disordered residues were mapped on the homology models of filamin domains. Different segments in each domain were numbered according to the schemes as shown in the figure 14. The topology diagram of filamin isoforms of CH domain consists of four main α-helices (1, 2, 3, 4) connected by loop regions (0-1, 1-2, 2-3, 3-4) that might contain two or three shorter helices. Three dominant helices form a parallel bundle against which N-terminal helix packs at a right angle. Similarly the topology diagram of Ig-like domain presents an immunoglobulin-like fold made up of seven β-strands organized in two beta sheets giving a β-sandwich. The first β -sheet consists of strands 1, 2, 5 and 6. The second β -sheet consists of strands 3, 4, 7 and 8. Only in some of filamin isoforms strand 4 is present. The loops connecting different beta-strand are 0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7 and 8-7.

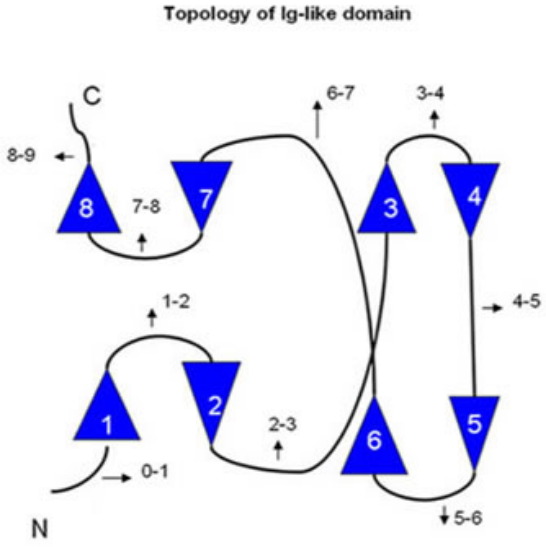
It can be seen that disordered residues were predicted for CH1 in the loop region 2-3 and for CH2 domain between the helix 2 and loop 3-4 and for Ig-like domains of all three human filamin isoforms mostly in the loop region (mostly in the loop between 1 and 2, and between 3 and 4). Moreover, the disordered residues are more often predicted at the N-terminus than at the C-terminus of the filamin protein (see table 14). For e.g. as shown in figure 15, predicted disordered residues were marked on homology models of Ig-like

domain 3.From the topology diagram, we can infer that disordered residues tend to be in the loop region (between loop 1 and loop 2) and at the N-terminus.

Among filamin protein, Ig-like domain 15 of filamin A and B are predicted to be ordered, whereas Ig-like domain 15 of filamin C is predicted to be disordered. Also Ig-like domain 24 of filamin A and C predicted to be completely ordered, whereas Ig-like domain 24 of filamin B is predicted to be partially disordered (see table 13). The overall fraction of disordered residues predicted in filamin domain is in range from 2% to 20%, in line with what is commonly observed in globular proteins (see section 4.5.8).



(i)



(ii)

Figure 14. (i) Toplogy diagram of the CH domain. α-helices (1, 2, 3, 4) connected by loop regions (0-1, 1-2, 2-3, 3-4). Three helices (2, 3, and 4) form a parallel bundle against which N-terminal helix (1) packs at a right angle. **(ii)** Topology diagram of the Ig-like domain. β-strands are organized in two beta sheets giving a β-sandwich fold. The first β -sheet consists of strands 1, 2, 5 and 6. The second β -sheet consists of strands 3, 4, 7 and 8. Strand 4 is present only in some domains. The loops connecting different beta-strands are 0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7 and 8-7.

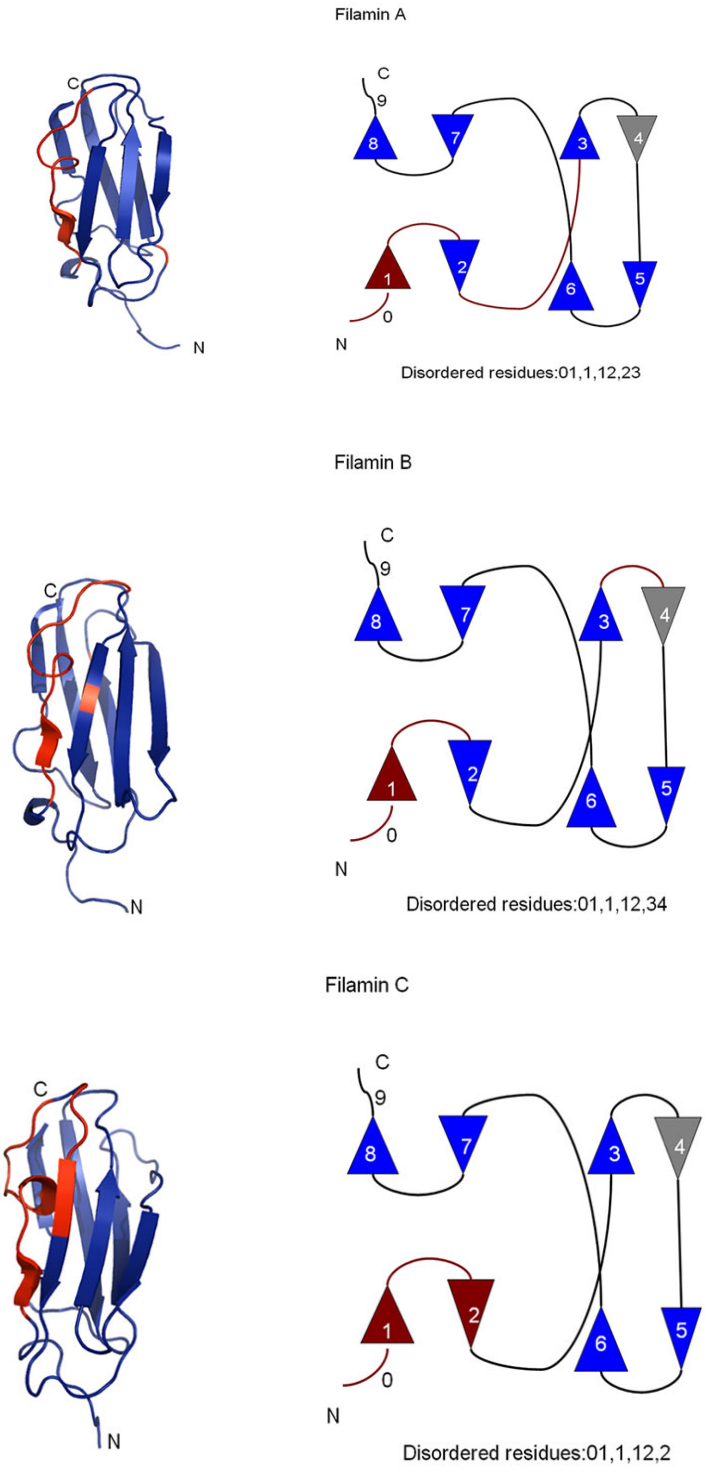


Figure 15. Disordered residues mapped on the homology models of Ig-like domain 3 of the three filamin isoforms. Blue regions indicate ordered residues, while disordered moieties are indicated in red. Strand 4 is not observed in this domain and is indicated in grey.

Table 13. Prediction of conformationally disordered residues mapped on the homology models. CH1_A, CH1_B, CH1_C, CH2_A, CH2_B and CH2_C are the CH domains (1 and 2) of the three filamin isoforms (A, B, and C). Within each CH domain there are four helices (1, 2, 3, and 4) and the loops between them (0-1, 1-2, 2-3, and 3-4). The 24 Ig-like domains are indicated as R1A, …, R24C. Each of them is a β-sandwich of two β-sheets. The eight β-strands are numbered from 1 to 8 and the loops between them are indicated as 0-1, 1-2, 2-3, 3-4, 4-5, 5-6, 6-7 and 8-7. See Figure 14 for details about the topology of these structural domains.

Domain	*01*	*1*	*12*	*2*	*23*	*3*	*34*	*4*
Ch1_A					X			
Ch1_B					X			
Ch1_C								
Ch2_A				X	X	X		
Ch2_B				X	X	X	X	
Ch2_C				-		X	X	-

Domain	*01*	*1*	*12*	*2*	*23*	*3*	*34*	*4*	*45*	*5*	*56*	*6*	*67*	*7*	*78*	*8*	*89*
R1A			X				X	-									
R1B			X	X	X	X	X	-									
R1C			X					-									
R2A			X					-		-	X						
R2B								-									
R2C			X					-		-		X					
R3A	X	X	X		X			-									
R3B	X	X	X				X	-									
R3C	X	X		X				-									
R4A							X	-			X	X	X				
R4B							X	-			X	X	X	X	X		
R4C								-			X	X	X	X	X		
R5A	X		X			X	X	-									
R5B			X					-									
R5C								-									
R6A						X	X	-									
R6B			X		X	X	X	X		-							
R6C			X				X	-									
R7A			X	X				-						X	X		
R7B			X	X	X	X	X	-									
R7C			X					-									
R8A							X	-	X	X	X						
R8B																	
R8C			X					-		X	X						

R9A			X		X			-									
R9B			X			X		-			X						
R9C			X		X	X		-									
R10A																	
R10B			X			X											
R10C																	
R11A							X	-		X	X	X	X	X	X		
R11B					X		X	-									
R11C			X				X	-									
R12A					X			-									
R12B					X	X	X	-			X						
R12C			X														
R13A			X					-									
R13B			X					-	X	X	X	X	X				
R13C		-	X				X	-		X	X						
R14A			X										X				
R14B			X					-					X	X	X		
R14C			X					-					X	X	X		
R15A		-						-									
R15B		-						-									
R15C		-				X	X	-									
R16A		-									X	X					
R16B		-															
R16C		-						-			X	X	X	X	X		
R17A								-		-	X						
R17B								-		-							
R17C			X					-			X						
R18A			-			X	X	-									
R18B			-				X	-									
R18C			-				X	-									
R19A					X	X	X	-				X					
R19B					X			-									
R19C								-									
R20A		-			X												
R20B		-			X		X	X									
R20C	X	X	X	-	X				X								
R21A								-			X	X	X	X	X		
R21B					X		X	-			X	X	X	X	X		
R21C								-			X	X	X	X	X		
R22A																	
R22B			X				X	-		X							
R22C		-						-		-							
R23A								-									
R23B						X		-									

R23C			X					-									
R24A								-									
R24B							X	-		X							
R24C								-									

X – Disordered residues predicted
- Not present
R1A,B,C-R24A,B,C-Filamin Ig-like domain 1-24

Table 14. Frequency of occurrence of conformational disorder in the 24 segments of the Ig-like domains of human filamin (percentage).

Segment	***Filamin Isoforms***			
	All	A	B	C
N-terminus	7	8	4	8
Strand 1	6	5	5	10
Loop 1-2	45	35	43	57
Strand 2	6	4	8	4
Loop 2-3	21	21	33	8
Strand 3	18	17	29	8
Loop 3-4	35	33	50	21
Strand 4	15	0	40	0
Loop 4-5	4	4	4	4
Strand 5	11	9	14	9
Loop 5-6	25	29	21	25
Strand 6	17	21	12	17
Loop 6-7	17	17	17	17
Strand 7	14	12	12	17
Loop 7-8	14	12	12	17
Strand 8	0	0	0	0
C-terminus	0	0	0	0

6.3.1.1. Conformational disorder in the experimental filamin PDB entries

We studied the conformation disorder status in experimentally solved structures which are deposited in the PDB. We inferred the conformational disordered status from the missing residues in the REMARK 465 entries. In filamin A Ig-like domain 17 along with GPIB alpha cytoplasmic domain complex (PDB ID: 2BP3), the disordered residues is in the loop 0-1 and in the loop 8-9. In filamin A Ig-like domain 21 complexed with MIG FILIN peptide (PDB ID: 2WOP), the conformation disordered is in the loop 0-1, loop 5-6 and loop 8-9. In filamin C Ig-like domain 23 (PDB ID: 2NOC), the conformational disordered residues is in the loop 0-1. The conformational disorder residues are only observed either in the N-terminus (loop 0-1) or at the C-terminus (loop 8-9) of Ig-like domain of human filamin. Despite the paucity of the data, they agree quite well with the predictions described above.

6.3.2. Quaternary status prediction of filamin protein

A set of protein chains containing monomeric, homo-oligomeric and hetero-oligomeric serve as training set (summarized in section 5.2.2). The 26 individual filamin domains are represented with their amino acid composition. As summarized in section 6.2.1, we tried several learning schemes on the training data and used a 10-fold cross validation to select the best performing algorithm. The classifier function-SMO within the WEKA package (http://sourceforge.net/projects/weka/) was selected. The quaternary status of the filamin A, filamin B, and filamin C of domains is predicted against the three types of learning sets, one with proteins containing less than 100 residues, one with 100-200 residue proteins, and the third with 200-300 proteins.

According to the expectation, all isoforms of filamin were predicted to be systematically homo-oligomeric, with very few expectations localized around the Ig-like domain 10-11 of filamin A and 19-20 of filamin B, which show a modest tendency to hetero-oligomerization.

6.3.3. Metal binding prediction of filamin protein

Predictions were done by using the random forest machine learning method, using the freely available package WEKA (http://sourceforge.net/projects/weka/). Learning sets and queries were prepared as described in section 5.3. Predictions were done on single filamin domains and on increasingly longer constructs by including, one by one, successive domains.

6.3.3.1. Metal binding prediction on individual filamin domains

It is observed that copper ion is predicted to be suitable to bind many domains of the filamin protein followed by cobalt ion which show a more modest tendency to be complexed by filamin domains. The least occurring metal ions are nickel, manganese and magnesium (table 21).

In filamin A the CH1 domain and Ig-like domain 24 have least metal presence. Metal affinity is predicted to be higher in Ig-like domain 15 followed by Ig-like domain 5, Ig-like domain 6 and Ig-like domain 7. In all the domains of filamin A, the most predicted metal ions are copper, cobalt, calcium, and zinc ion (table 15).

In filamin B, Ig-like domain 13, and Ig-like domain 15 show the absence of metal ion. The least presence of metal ion is predicted to be in CH1 domain, CH2 domain and Ig-like domain 24. The major occurrence of metal ion is predicted among Ig-like domain 2, Ig-like domain 6, Ig-like domain 11 and Ig-like domain 21.Among all the domains, the most predicted metal ions are copper, iron, cobalt, calcium and zinc and manganese, magnesium and nickel are the least predicted (table 16).

In filamin C, the CH2 domain, Ig-like domain 3, Ig-like domain 8 shows the absence of metal ion. The major occurrence of metal ion is predicted among the domains Ig-like domain 1, Ig-like domain 2. The least occurrence of metal ion is predicted among the domains CH1 domain, Ig-like domain 18, Ig-like domain 19 and Ig-like domain 24. Among all the domains, the most predicted metal ions are copper, cobalt, calcium and iron and the least predicted metal ions are nickel, magnesium, manganese and zinc (table 17).

6.3.3.2. Metal binding prediction on large filamin segments

In large segments of filamin protein, cobalt ion is the most frequently cation is predicted to be complexed by the protein followed by copper ion. The least occurring metal ions are magnesium, nickel and zinc (table 22).

In filamin A, the most frequently predicted cations are copper and cobalt followed by iron and calcium. The least frequently predicted metal ions are manganese and zinc. The metal ions nickel and magnesium is completely absent in predictions. The large segments

like CH1-CH2-R1-R18, CH1-CH2-R1-R19 are predicted to contain more metal ions and least metal occurrence is predicted in the segments CH1-CH2, CH1-CH2-R1. The metal ions calcium, cobalt and copper are dominant in segments CH1-CH2-R1-R16 to CH1-CH2-R1-R24 (table 18).

In filamin B, the most frequently predicted metal ions are copper, calcium, and iron and the least frequently predicted metal ions are nickel and zinc. The magnesium ion is predicted to be completely absent in all segments. Between the large segments more ions are predicted in CH1-CH1, CH1-CH2-R1 and CH1-CH2-R1-R2 and less metal ions are predicted in CH1-CH2-R1-R20 and CH1-CH2-R1-R21. It is also observed that occurrence of metal ion is predicted to decrease as the segments gets larger (table 19).

In filamin C, cobalt and copper ions are frequently predicted and the least frequently predicted metal ions are magnesium, iron, and zinc. The cobalt ion is predicted in all the segments except in the segment CH1-CH2. The copper ion is dominant starting from the segment CH1-CH2-R1-R5 to CH1-CH2-R1-R24. Between the segments, more metal ions are predicted for the segment CH1-CH2 to CH1-CH2-R1-R8 and the least occurrence metal ions is predicted in CH1-CH2-R1-R9. The presence of metal ions like magnesium, manganese, nickel and zinc are predicted only between segments CH1-CH2 to the segment CH1-CH2-R1-R11. From the segments CH1-CH2-R1-R15 to CH1-CH2-R1-R24 metal ions like cobalt, copper and iron are dominant (table 20).

Table 15. Metal binding prediction in individual domains of filamin A. CH1 and CH2 represents the CH domains. R1, R2...R23, R24 represent the Ig-like domains. In this table, 'M' indicated that the metal ion is predicted to be present and the sign '-' indicates the absence of metal ion.

	Ca	***Co***	***Cu***	***Fe***	***Mg***	***Mn***	***Ni***	***Zn***
CH1	-	-	-	-	-	-	M	-
CH2	M	-	M	-	-	-		-
R1	-	M	-	-	-	-	M	-
R2	-	-	M	-	-	-	-	-
R3	-	-	-	-	-	-	-	-
R4	-	M	-	-	M	M	-	M
R5	-	M	M	-	-	-	-	M
R6	M	M	M	-	-	-	-	-
R7	-	M	M	-	-	-	-	M
R8	-	M	-	-	-	-	-	M
R9	-	-	M	-	-	-	-	-
R10	M	-	M	-	-	-	-	-
R11	M	-	M	-	-	-	-	-
R12	-	-	M	-	-	-	-	M
R13	-	-	M	-	-	M	-	-
R14	-	-	M	-	-	-	-	-
R15	-	M	M	-	M	M	-	-
R16	M	M	M	-	-	-	-	-
R17	M	-	M	-	-	-	-	-
R18	-	M	-	-	-	-	-	-
R19	M	-	M	-	-	-	-	-
R20	-	-	-	-	-	-	-	-
R21	-	M	-	-	-	-	-	M
R22	-	-	-	-	-	-	-	-
R23	M	-	M	-	-	-	-	-
R24	-	M	-	-	-	-	-	-

Table 16. Metal binding prediction in individual domains of filamin B. CH1 and CH2 represents the CH domains. R1, R2...R23, R24 represent the Ig-like domains. In this table, 'M' indicated that the metal ion is predicted to be present and the sign '-' indicates the absence of metal ion.

	Ca	***Co***	***Cu***	***Fe***	***Mg***	***Mn***	***Ni***	***Zn***
CH1	-	-	-	-	-	-	M	-
CH2	-	-	-	M	-	-	-	-
R1	-	M	-	-	-	-	-	M
R2	M	-	M	M	-	-	M	-
R3	-	-	M	-	-	-	-	M
R4	M	-	M	M	-	-	-	-
R5	M	M	M	-	-	-	-	-
R6	M	M	M	M	-	M	-	-
R7	M	-	M	-	-	-	-	M
R8	-	M	M	-	M	-	-	-
R9	-	-	M	-	-	-	-	-
R10	-	-	M	-	-	-	-	M
R11	M	-	M	M	-	M	-	M
R12	-	-	M	-	-	-	-	M
R13	-	-	-	-	-	-	-	-
R14	M	-	M	-	-	-	-	-
R15	-	-	-	-	-	-	-	-
R16	-	-	M	M	-	-	-	M
R17	M	-	M	-	-	-	-	-
R18	-	-	M	-	-	-	-	-
R19	-	-	M	-	-	-	-	M
R20	-	M	M	-	-	M	-	-
R21	-	-	-	M	-	-	-	M
R22	-	M	M	-	-	-	-	-
R23	-	-	M	-	-	-	-	-
R24	-	M	-	-	-	-	-	-

Table 17. Metal binding prediction in individual domains of filamin C. CH1 and CH2 represents the CH domains. R1, R2...R23, R24 represent the Ig-like domains. In this table, 'M' indicated that the metal ion is predicted to be present and the sign '-' indicates the absence of metal ion.

	Ca	***Co***	***Cu***	***Fe***	***Mg***	***Mn***	***Ni***	***Zn***
CH1	-	-	-	-	-	-	M	-
CH2	-	-	-	-	-	-	-	-
R1	-	M	M	M	M	-	M	-
R2	-	M	M	M	-	-	-	-
R3	-	-	-	-	-	-	-	-
R4	-	M	M	M	-	M	-	M
R5	-	M	-	-	-	-	-	M
R6	M	M	M	-	-	-	-	M
R7	M	M	M	-	-	-	-	-
R8	-	-	-	-	-	-	-	-
R9	-	-	M	-	-	-	-	-
R10	-	-	M	-	-	-	-	-
R11	M	M	M	-	-	-	-	-
R12	-	-	M	-	-	-	-	-
R13	-	M	M	-	-	-	-	-
R14	M	-	M	-	-	-	-	M
R15	-	M	M	-	-	-	M	-
R16	-	M	-	M	-	-	-	-
R17	-	-	M	-	-	-	-	-
R18	-	-	-	-	-	M	-	-
R19	-	-	-	-	M	-	-	-
R20	M	-	M	M	-	-	-	-
R21	-	M	-	-	M	-	-	M
R22	-	-	M	-	-	M	-	M
R23	-	-	M	-	-	-	-	M
R24	-	-	-	-	-	-	-	M

Table 18. Metal binding prediction in large segments of filamin A. Each segment contains an increasing portion of the protein. For example, CH1-CH2 is a construct that consists of the first two CH domains. CH-R1 includes also the first Ig-like domain besides the two CH domains. The construct is enlarged until CH-R1-R24, which contains the entire protein. In this table, 'M' represents metal ion predicted and '-' represents absence of metal ion.

	Ca	*Co*	*Cu*	*Fe*	*Mg*	*Mn*	*Ni*	*Zn*
CH1-CH2	-	M	-	-	-	-	-	M
CH-R1	-	M	-	-	-	-	-	-
CH- R1-R2	M	M	-	-	-	-	-	M
CH- R1-R3	-	M	-	-	-	-	-	-
CH- R1-R4	-	M	-	-	-	-	-	-
CH- R1-R5	-	M	-	-	-	-	-	-
CH- R1-R6	-	M	-	-	-	-	-	-
CH- R1-R7	-	M	M	-	-	-	-	-
CH- R1-R8	-	M	M	-	-	-	-	-
CH- R1-R9	-	M	M	-	-	-	-	-
CH- R1-R10	-	M	M	M	-	-	-	-
CH- R1-R11	-	M	M	M	-	-	-	-
CH- R1-R12	-	M	M	-	-	-	-	-
CH- R1-R13	-	M	M	-	-	-	-	-
CH- R1-R14	-	M	M	-	-	-	-	-
CH- R1-R15	-	M	M	-	-	-	-	-
CH- R1-R16	M	M	M	-	-	-	-	-
CH- R1-R17	M	M	M	-	-	-	-	-
CH- R1-R18	M	M	M	-	-	M	-	M
CH- R1-R19	M	M	M	-	-	M	-	M
CH- R1-R20	M	M	M	-	-	-	-	M
CH- R1-R21	M	M	M	M	-	-	-	-
CH- R1-R22	M	M	M	M	-	-	-	-
CH- R1-R23	M	M	M	-	-	-	-	-
CH- R1-R24	-	M	M	M	-	-	-	M

Table 19. Metal binding prediction in large segments of filamin B. Each segment contains an increasing portion of the protein. For example, CH1-CH2 is a construct that consists of the first two CH domains. CH-R1 includes also the first Ig-like domain besides the two CH domains. The construct is enlarged until CH-R1-R24, which contains the entire protein. In this table, 'M' represents metal ion predicted and '-' represents absence of metal ion.

	Ca	***Co***	***Cu***	***Fe***	***Mg***	***Mn***	***Ni***	***Zn***
CH1-CH2	-	-	-	M	-	M	M	M
CH-R1	M	-	-	M	-	M	-	M
CH- R1-R2	M	M	-	-	-	M	-	M
CH- R1-R3	-	M	-	-	-	M	-	M
CH- R1-R4	M	-	-	-	-	-	-	M
CH- R1-R5	M	M	-	-	-	M	-	M
CH- R1-R6	M	M	-	-	-	M	-	-
CH- R1-R7	M	M	M	-	-	-	-	-
CH- R1-R8	M	M	M	-	-	-	-	-
CH- R1-R9	M	M	M	-	-	-	-	-
CH- R1-R10	-	M	M	-	-	-	-	-
CH- R1-R11	M	M	M	-	-	-	-	-
CH- R1-R12	M	M	M	-	-	-	-	-
CH- R1-R13	-	M	M	M	-	-	-	-
CH- R1-R14	-	M	M	M	-	-	-	-
CH- R1-R15	-	M	M	M	-	-	-	-
CH- R1-R16	-	M	M	M	-	-	-	-
CH- R1-R17	-	M	M	M	-	-	-	-
CH- R1-R18	-	M	M	M	-	-	-	-
CH- R1-R19	-	M	M	M	-	-	-	-
CH- R1-R20	-	M	M	-	-	-	-	-
CH- R1-R21	-	-	M	-	-	-	-	-
CH- R1-R22	-	-	M	-	-	-	-	-
CH- R1-R23	-	M	M	-	-	-	-	-
CH- R1-R24	-	M	M	-	-	-	-	-

Table 20. Metal binding prediction in large segments of filamin C. Each segment contains an increasing portion of the protein. For example, CH1-CH2 is a construct that consists of the first two CH domains. CH-R1 includes also the first Ig-like domain besides the two CH domains. The construct is enlarged until CH-R1-R24, which contains the entire protein. In this table, 'M' represents metal ion predicted and '-' represents absence of metal ion.

	Ca	*Co*	*Cu*	*Fe*	*Mg*	*Mn*	*Ni*	*Zn*
CH1-CH2	-	-	-	M	-	M	M	M
CH-R1	-	M	-	M	-	M	M	M
CH- R1-R2	-	M	-	M	-	M	M	-
CH- R1-R3	-	M	-	M	M	M	M	-
CH- R1-R4	-	M	-	M	M	M	M	-
CH- R1-R5	-	M	M	-	-	M	M	M
CH- R1-R6	-	M	M	-	-	M	M	M
CH- R1-R7	-	M	M	-	-	-	M	-
CH- R1-R8	-	M	M	-	-	-	M	-
CH- R1-R9	-	M	-	-	-	M	-	-
CH- R1-R10	-	M	M	-	-	-	-	-
CH- R1-R11	-	M	M	-	-	M	-	-
CH- R1-R12	-	M	M	-	-	M	-	-
CH- R1-R13	-	M	M	-	-	-	-	-
CH- R1-R14	-	M	M	-	-	-	-	-
CH- R1-R15	-	M	M	M	-	-	-	-
CH- R1-R16	-	M	M	M	-	-	-	-
CH- R1-R17	-	M	M	M	-	-	-	-
CH- R1-R18	-	M	M	M	-	-	-	-
CH- R1-R19	-	M	M	M	-	-	-	-
CH- R1-R20	-	M	M	M	-	-	-	-
CH- R1-R21	-	M	M	M	-	-	-	-
CH- R1-R22	-	M	M	M	-	-	-	-
CH- R1-R23	-	M	M	M	-	-	-	-
CH- R1-R24	M	M	M	M	-	-	-	-

It can be concluded that filamin shows a considerable tendency to uptake metal cations.

The physiological role of these interactions, however, needs an experimental validation.

Table 21. Summary of the metal binding prediction in individual domains of filamin A, B, and C. CH1 and CH2 represents the CH domains. R1, R2...R23, R24 represent the Ig-like domains. In this table, 'M' indicated that the metal ion is predicted to be present and the sign '-' indicates the absence of metal ion.

	Ca			***Co***			***Cu***			***Fe***			***Mg***			***Mn***			***Ni***			***Zn***		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
CH1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M	M	M	-	-	-
CH2	M	-	-	-	-	-	M	-	-	-	M	-	-	-	-	-	-	-		-	-	-	-	-
R1	-	-	-	M	M	M	-	-	M	-	-	M	-	-	M	-	-	-	M	-	M	-	M	-
R2	-	M	-	-	-	M	M	M	M	-	M	M	-	-	-	-	-	-	-	M	-	-	-	-
R3	-	-	-	-	-	-	-	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M	-
R4	-	M	-	M	-	M	-	M	M	-	M	M	M	-	-	M	-	M	-	-	-	M	-	M
R5	-	M	-	M	M	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	M	-	M
R6	M	M	M	M	M	M	M	M	M	-	M	-	-	-	-	-	M	-	-	-	-	-	-	M
R7	-	M	M	M	-	M	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	M	M	-
R8	-	-	-	M	M	-	-	M	-	-	-	-	-	M	-	-	-	-	-	-	-	M	-	-
R9	-	-	-	-	-	-	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R10	M	-	-	-	-	-	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	M	-
R11	M	M	M	-	-	M	M	M	M	-	M	-	-	-	-	-	M	-	-	-	-	-	M	-
R12	-	-	-	-	-	-	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	M	M	-
R13	-	-	-	-	-	M	M	-	M	-	-	-	-	-	-	M	-	-	-	-	-	-	-	-
R14	-	M	M	-	-	-	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M
R15	-	-	-	M	-	M	M	-	M	-	-	-	M	-	-	M	-	-	-	-	M	-	-	-
R16	M	-	-	M	-	M	M	M	-	-	M	M	-	-	-	-	-	-	-	-	-	-	M	-
R17	M	M	-	-	-	-	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R18	-	-	-	M	-	-	-	M	-	-	-	-	-	-	-	-	-	M	-	-	-	-	-	-
R19	M	-	-	-	-	-	M	M	-	-	-	-	-	-	M	-	-	-	-	-	-	-	M	-
R20	-	-	M	-	M	-	-	M	M	-	-	M	-	-	-	-	M	-	-	-	-	-	-	-
R21	-	-	-	M	-	M	-	-	-	-	M	-	-	-	M	-	-	-	-	-	-	M	M	M
R22	-	-	-	-	M	-	-	M	M	-	-	-	-	-	-	-	-	M	-	-	-	-	-	M
R23	M	-	-	-	-	-	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M
R24	-	-	-	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M

Table 22. Summary of the metal binding prediction in large segments of filamin A, B, and C. Each segment contains an increasing portion of the protein. For example, CH1-CH2 is a construct that consists of the first two CH domains. CH-R1 includes also the first Ig-like domain besides the two CH domains. The construct is enlarged until CH-R1-R24, which contains the entire protein. In this table, 'M' represents metal ion predicted and '-' represents absence of metal ion.

Segment	***Ca***			***Co***			***Cu***			***Fe***			***Mg***			***Mn***			***Ni***			***Zn***		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
CH1-CH2	-	-	-	M	-	-	-	-	-	-	M	M	-	-	-	-	M	M	-	M	M	M	M	M
CH-R1	-	M	-	M	-	M	-	-	-	-	M	M	-	-	-	-	M	M	-	-	M	-	M	M
CH- R1-R2	M	M	-	M	M	M	-	-	-	-	-	M	-	-	-	-	M	M	-	-	M	M	M	-
CH- R1-R3	-	-	-	M	M	M	-	-	-	-	-	M	-	-	M	-	M	M	-	-	M	-	M	-
CH- R1-R4	-	M	-	M	-	M	-	-	-	-	-	M	-	-	M	-	-	M	-	-	M	-	M	-
CH- R1-R5	-	M	-	M	M	M	-	-	M	-	-	-	-	-	-	-	M	M	-	-	M	-	M	M
CH- R1-R6	-	M	-	M	M	M	-	-	M	-	-	-	-	-	-	-	M	M	-	-	M	-	-	M
CH- R1-R7	-	M	-	M	M	M	M	M	M	-	-	-	-	-	-	-	-	-	-	-	M	-	-	-
CH- R1-R8	-	M	-	M	M	M	M	M	M	-	-	-	-	-	-	-	-	-	-	-	M	-	-	-
CH- R1-R9	-	M	-	M	M	M	M	M	-	-	-	-	-	-	-	-	-	M	-	-	-	-	-	-
CH- R1-R10	-	-	-	M	M	M	M	M	M	M	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R11	-	M	-	M	M	M	M	M	M	M	-	-	-	-	-	-	-	M	-	-	-	-	-	-
CH- R1-R12	-	M	-	M	M	M	M	M	M	-	-	-	-	-	-	-	-	M	-	-	-	-	-	-
CH- R1-R13	-	-	-	M	M	M	M	M	M	-	M	-	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R14	-	-	-	M	M	M	M	M	M	-	M	-	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R15	-	-	-	M	M	M	M	M	M	-	M	M	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R16	M	-	-	M	M	M	M	M	M	-	M	M	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R17	M	-	-	M	M	M	M	M	M	-	M	M	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R18	M	-	-	M	M	M	M	M	M	-	M	M	-	-	-	M	-	-	-	-	-	M	-	-
CH- R1-R19	M	-	-	M	M	M	M	M	M	-	M	M	-	-	-	M	-	-	-	-	-	M	-	-
CH- R1-R20	M	-	-	M	M	M	M	M	M	-	-	M	-	-	-	-	-	-	-	-	-	M	-	-
CH- R1-R21	M	-	-	M	-	M	M	M	M	M	-	M	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R22	M	-	-	M	-	M	M	M	M	M	-	M	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R23	M	-	-	M	M	M	M	M	M	-	-	M	-	-	-	-	-	-	-	-	-	-	-	-
CH- R1-R24	-	-	M	M	M	M	M	M	M	M	-	M	-	-	-	-	-	-	-	-	-	M	-	-

6.4. Protein domain boundary predictions

Here we summarize a study that was published (Kirillova et al. 2009), and which is attached (section 8.2). In structural biology, computational tools for predicting domain boundaries play a vital role to design of protein constructs that must be expressed in a stable and functional form. However, prediction of protein domain boundaries on the basis of amino acid sequence is still very problematical. In this published work, the performance of several computational approaches that were made publicly available in CASP 7 experiment is compared and the reliability of these prediction methods for practical application in structural biology was tested.

In CASP experiments, the three-dimensional structures of protein sequences which were determined experimentally though they were not yet published are distributed to participants. The key feature is that participants make blind predictions and these predictions are assessed in comparison to the reality.

For prediction of domain boundaries, the data were obtained from the CASP7 web page (http://predictioncenter.gc.ucdavis.edu/casp7/), for which both predictions and experimental data are publicly available. The bioinformatics tools that are freely available in CASP7 were examined (table 23). Given the ambiguity in protein domain definition, the real boundaries were defined according to the CASP7 organizers and assessors.

Table 23. Publicly bioinformatics tools used in CASP7 to predict domain boundaries.

Tools	***URL***
Baker	http://robetta.org/submit.jsp
Chop	http://cubic.bioc.columbia.edu/services/chop/index.htm
Chophomo	http://www.cubic.bioc.columbia.edu/services/chop/index.htm
Distill	http://distill.ucd.ie/distill
Domfold	http://ww.reading.ac.uk/bioinf/Domfold
Domssea	http://bioinf.cs.ucl.ac.uk/dompred
Dps	http://bioinf.cs.ucl.ac.uk/dompred
Foldpro	http://www.igb.uci.edu/servers/psss.html
Hhpred1	http://toolkit.tuebingen.mpg.de/hhpred
Hhpred 3	http://toolkit.tuebingen.mpg.de/hhpred
Maopus	http://sigler.bioch.bcm.tcm.edu/CASP7-DOM
Metadp	http://meta-dp.cse.buffalo.edu
NNput	http://webmobis.cs.put.poznan.pl/webmobis/app
Robetta	http://robetta.org/submit.jsp

To predict, on the basis of the protein length, that a protein contains one domain or it is multi-domain, a threshold value t was used. Only proteins shorter than t were considered. Table 24 show the values of the Matthews correlation coefficient (MCC) (see methods-section 8.2) observed at various threshold values for the proteins examined in the CASP7 experiment. The highest MCC (0.628) is observed at t=200 residues. This prediction strategy is compared with the Matthews correlation coefficient values computed on the basis of the predictions deposited by the participants to the CASP7. It is observed that most of the bioinformatics tools in CASP7 are less reliable than the predictions based on the very simple assumption that a small protein has a high probability to contain a single domain and that a large protein is likely to contain two or more domains.

Predicted and real partitions were compared with the Jaccard index, the Rand coefficient and the Fowlkes-Mallows index and their statistical significance were calculated (see

methods-section 8.2).Based on this statistical calculation it was observed that matching between prediction and reality is slightly better for small protein than for large proteins.

Table 24. Matthews Correlation coefficients (MCC) at various threshold values (t).

T	*MCC*
70	0.063
80	0.111
90	0.173
100	0.233
110	0.276
120	0.307
130	0.367
140	0.397
150	0.469
160	0.535
170	0.582
180	0.586
190	0.614
200	0.628
210	0.544
220	0.559
230	0.510
240	0.445
250	0.462
260	0.346
270	0.330

The accuracy with which the domain boundaries are identified by various prediction methods was examined. Table 25 show the percentage of domains that are correctly predicted and the difference between the real and the predicted boundary in the subset of domains that are correctly predicted. It is observed that the percentage of good

predictions is about 30-40%, though some of the prediction methods are better than others and correctly identify about 60% of the domains. When predictions are good, however, they are really excellent. Both the N- and C-terminal boundaries are identified with high accuracy-the average Delta_b and Delta_e values are close to 0.

Table 25. Accuracy with which the domain boundaries are identified by various prediction methods. Delta_b is the difference between the sequence position in which the domain is predicted to begin and the sequence position in which it begins in the reality. Delta_e is the difference between the sequence position in which the domain is predicted to end and the sequence position in which it ends in the reality. Pc_c is the average deviation between the real and the predicted beginning of the domain Delta_b and Delta_e are the average difference between the real and the predicted end of the domain (standard deviations of the mean in the parentheses).

Method	***Pc_c***	***Delta_b***	***Delta_e***
baker	56.2	-1.20(0.3)	2.2(0.5)
Chop	26.1	-2.9(1.0)	1.9(0.7)
Chophomo	25.0	-2.6(1.0)	2.9(1.0)
Distill	33.6	-1.5(0.6)	3.2(0.8)
Domfold	38.0	-1.9(0.6)	2.9(0.7)
Domssea	42.9	-1.7(0.6)	2.5(0.7)
Dps	38.7	-2.2(0.8)	1.6(0.9)
Foldpro	62.8	-1.30(0.4)	2.0(0.4)
Hhpred1	43.3	-2.1(0.5	2.6(0.5)
Hhpred3	43.4	-2.1(0.5)	2.7(0.5)
Maopus	54.2	-1.4(0.6)	3.0(0.8)
Metadp	39.8	-1.3(0.7)	3.3(0.7)
NNput	30.8	-1.90(0.7)	2.4(0.8)
Robetta	57.9	-1.0(0.3)	1.5(0.5)

It can be concluded that bioinformatics tools are still immature and are not yet sufficiently accurate to be used as routine tools in experimental structural biology, though some of them are rather promising.

7. References

1. Andreini, C., Bertini, I., and Rosato, A. 2004. A hint to search for metalloproteins in gene banks. *Bioinformatics* **20:** 1373-1380.
2. Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181:** 223-230.
3. Arron, J.R., Pewzner-Jung, Y., Walsh, M.C., Kobayashi, T., and Choi, Y. 2002. Regulation of the subcellular localization of tumor necrosis factor receptor-associated factor (TRAF)2 by TRAF1 reveals mechanisms of TRAF2 signaling. *J Exp Med* **196:** 923-934.
4. Awata, H., Huang, C., Handlogten, M.E., and Miller, R.T. 2001. Interaction of the calcium-sensing receptor and filamin, a potential scaffolding protein. *J Biol Chem* **276:** 34871-34879.
5. Babor, M., Gerzon, S., Raveh, B., Sobolev, V., and Edelman, M. 2008. Prediction of transition metal-binding sites from apo protein structures. *Proteins* **70:** 208-217.
6. Berry, F.B., O'Neill, M.A., Coca-Prados, M., and Walter, M.A. 2005. FOXC1 transcriptional regulatory activity is impaired by PBX1 in a filamin A-mediated manner. *Mol Cell Biol* **25:** 1415-1424.
7. Burley, S.K., Joachimiak, A., Montelione, G.T., and Wilson, I.A. 2008. Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. *Structure* **16:** 5-11.
8. Byvatov, E., and Schneider, G. 2003. Support vector machine applications in bioinformatics. *Appl Bioinformatics* **2:** 67-77.
9. Carugo,O., Djinovic Carugo, K., Gorbalenya, A.E., & Tucker, P.2007a. Likelihood of crystallization: experimental and computational approaches.J.Appl.Cryst.40, 292-293.
10. Carugo,O.,Djinovic Carugo,K., Gorbalenya, A.E., & Tucker, P.2007b. Editorial on Hot Topic: Workshop on the definition of protein domains and their likelihood of crystallization.Curr.Prot.Pept.Sci.8, 119-120.
11. Carugo,O. 2007c. A structural proteomics filter:prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences.J.Appl.Cryst.40(6):986-989.
12. Carugo, O. 2008d. Amino acid composition and protein dimension. Protein Sci 17: 2187-2191.
13. Chakarova, C., Wehnert, M.S., Uhl, K., Sakthivel, S., Vosberg, H.P., van der Ven, P.F., and Furst, D.O. 2000. Genomic structure and fine mapping of the two human filamin gene paralogues FLNB and FLNC and comparative analysis of the filamin gene family. *Hum Genet* **107:** 597-611.
14. Chakrabarti, P., and Pal, D. 2001. The interrelationships of side-chain and main-chain conformations in proteins. *Prog Biophys Mol Biol* **76:** 1-102.
15. Chen, K., Kurgan, L., and Rahbari, M. 2007. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem Biophys Res Commun* **355:** 764-769.
16. Chen, L., Oughtred, R., Berman, H.M., and Westbrook, J. 2004. TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20:** 2860-2862.

17. Chen, M., and Stracher, A. 1989. In situ phosphorylation of platelet actin-binding protein by cAMP-dependent protein kinase stabilizes it against proteolysis by calpain. *J Biol Chem* **264:** 14282-14289.
18. Chou, K.C., and Cai, Y.D. 2003. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins* **53:** 282-289.
19. Coeytaux, K., and Poupon, A. 2005. Prediction of unfolded segments in a protein sequence based on amino acid composition. *Bioinformatics* **21:** 1891-1900.
20. Cranmer, S.L., Pikovski, I., Mangin, P., Thompson, P.E., Domagala, T., Frazzetto, M., Salem, H.H., and Jackson, S.P. 2005. Identification of a unique filamin A binding region within the cytoplasmic domain of glycoprotein Ibalpha. *Biochem J* **387:** 849-858.
21. Djinovic-Carugo, K., Gautel, M., Ylanne, J., and Young, P. 2002. The spectrin repeat: a structural platform for cytoskeletal protein assemblies. *FEBS Lett* **513:** 119-123.
22. Djinovic Carugo, K., Banuelos, S., and Saraste, M. 1997. Crystal structure of a calponin homology domain. *Nat Struct Biol* **4:** 175-179.
23. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21:** 3433-3434.
24. Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J.E. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput***:** 473-484.
25. Dunker, A.K., Lawson, J.D., Brown, C.J., Williams, R.M., Romero, P., Oh, J.S., Oldfield, C.J., Campen, A.M., Ratliff, C.M., Hipps, K.W., et al. 2001. Intrinsically disordered protein. *J Mol Graph Model* **19:** 26-59.
26. Dunker, A.K., and Obradovic, Z. 2001. The protein trinity--linking function and disorder. *Nat Biotechnol* **19:** 805-806.
27. Dyson, J.M., Munday, A.D., Kong, A.M., Huysmans, R.D., Matzaris, M., Layton, M.J., Nandurkar, H.H., Berndt, M.C., and Mitchell, C.A. 2003. SHIP-2 forms a tetrameric complex with filamin, actin, and GPIb-IX-V: localization of SHIP-2 to the activated platelet actin cytoskeleton. *Blood* **102:** 940-948.
28. Enz, R. 2002. The actin-binding protein Filamin-A interacts with the metabotropic glutamate receptor type 7. *FEBS Lett* **514:** 184-188.
29. Ernst, J., Beg, Q.K., Kay, K.A., Balazsi, G., Oltvai, Z.N., and Bar-Joseph, Z. 2008. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Comput Biol* **4:** e1000044.
30. Ferron, F., Longhi, S., Canard, B., and Karlin, D. 2006. A practical overview of protein disorder prediction methods. *Proteins* **65:** 1-14.
31. Fink, A.L. 2005. Natively unfolded proteins. *Curr Opin Struct Biol* **15:** 35-41.
32. Fiser, A., and Sali, A. 2003. ModLoop: automated modeling of loops in protein structures. *Bioinformatics* **19:** 2500-2501.
33. Garcia, E., Stracher, A., and Jay, D. 2006. Calcineurin dephosphorylates the C-terminal region of filamin in an important regulatory site: a possible mechanism for filamin mobilization and cell signaling. *Arch Biochem Biophys* **446:** 140-150.

34. Gardel, M.L., Nakamura, F., Hartwig, J.H., Crocker, J.C., Stossel, T.P., and Weitz, D.A. 2006. Prestressed F-actin networks cross-linked by hinged filamins replicate mechanical properties of cells. *Proc Natl Acad Sci U S A* **103:** 1762-1767.
35. Garian, R. 2001. Prediction of quaternary structure from primary structure. *Bioinformatics* **17:** 551-556.
36. Gil-Pita, R., and Yao, X. 2008. Evolving edited k-nearest neighbor classifiers. *Int J Neural Syst* **18:** 459-467.
37. Gimona, M., and Mital, R. 1998. The single CH domain of calponin is neither sufficient nor necessary for F-actin binding. *J Cell Sci* **111 (Pt 13):** 1813-1821.
38. Goldbaum, M.H. 2005. Unsupervised learning with independent component analysis can identify patterns of glaucomatous visual field defects. *Trans Am Ophthalmol Soc* **103:** 270-280.
39. Goldsmith, S.C., Pokala, N., Shen, W., Fedorov, A.A., Matsudaira, P., and Almo, S.C. 1997. The structure of an actin-crosslinking domain from human fimbrin. *Nat Struct Biol* **4:** 708-712.
40. Goodford, P.J. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* **28:** 849-857.
41. Gorlin, J.B., Yamin, R., Egan, S., Stewart, M., Stossel, T.P., Kwiatkowski, D.J., and Hartwig, J.H. 1990. Human endothelial actin-binding protein (ABP-280, nonmuscle filamin): a molecular leaf spring. *J Cell Biol* **111:** 1089-1105.
42. Gravante, B., Barbuti, A., Milanesi, R., Zappi, I., Viscomi, C., and DiFrancesco, D. 2004. Interaction of the pacemaker channel HCN1 with filamin A. *J Biol Chem* **279:** 43847-43853.
43. Hendrickson, W.A. 2007. Impact of structures from the protein structure initiative. *Structure* **15:** 1528-1529.
44. Huber, R. 1979. Conformational flexibility in protein molecules. *Nature* **280:** 538-539.
45. Joachimiak, A. 2009. High-throughput crystallography for structural genomics. *Curr Opin Struct Biol*.
46. Jones, S., and Thornton, J.M. 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93:** 13-20.
47. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., and MacCoss, M.J. 2007. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4:** 923-925.
48. Kirillova, S., Kumar, S., and Carugo, O. 2009. Protein domain boundary predictions: a structural biology perspective. *Open Biochem J* **3:** 1-8.
49. Klotz, I.M., Langerman, N.R., and Darnall, D.W. 1970. Quaternary structure of proteins. *Annu Rev Biochem* **39:** 25-62.
50. Kumar, S., and Carugo, O. 2008. Consensus prediction of protein conformational disorder from amino acidic sequence. *Open Biochem J* **2:** 1-5.
51. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafe, G., Perez, A., et al. 2006. Machine learning in bioinformatics. *Brief Bioinform* **7:** 86-112.

52. Levy, R., Edelman, M., and Sobolev, V. 2009. Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* **76:** 365-374.
53. Li, W., and Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22:** 1658-1659.
54. Lin, R., Karpa, K., Kabbani, N., Goldman-Rakic, P., and Levenson, R. 2001. Dopamine D2 and D3 receptors are linked to the actin cytoskeleton via interaction with filamin A. *Proc Natl Acad Sci U S A* **98:** 5258-5263.
55. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., and Russell, R.B. 2003a. Protein disorder prediction: implications for structural proteomics. *Structure* **11:** 1453-1459.
56. Linding, R., Russell, R.B., Neduva, V., and Gibson, T.J. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31:** 3701-3708.
57. Lippi, M., Passerini, A., Punta, M., Rost, B., and Frasconi, P. 2008. MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics* **24:** 2094-2095.
58. Liu, G., Thomas, L., Warren, R.A., Enns, C.A., Cunningham, C.C., Hartwig, J.H., and Thomas, G. 1997. Cytoskeletal protein ABP-280 directs the intracellular trafficking of furin and modulates proprotein processing in the endocytic pathway. *J Cell Biol* **139:** 1719-1733.
59. Loy, C.J., Sim, K.S., and Yong, E.L. 2003. Filamin-A fragment localizes to the nucleus to regulate androgen receptor and coactivator functions. *Proc Natl Acad Sci U S A* **100:** 4562-4567.
60. Marsden, R.L., Lewis, T.A., and Orengo, C.A. 2007. Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics* **8:** 86.
61. Marti, A., Luo, Z., Cunningham, C., Ohta, Y., Hartwig, J., Stossel, T.P., Kyriakis, J.M., and Avruch, J. 1997. Actin-binding protein-280 binds the stress-activated protein kinase (SAPK) activator SEK-1 and is required for tumor necrosis factor-alpha activation of SAPK in melanoma cells. *J Biol Chem* **272:** 2620-2628.
62. McPherson, A. 2004. Introduction to protein crystallization. *Methods* **34:** 254-265.
63. Melville, J.L., Burke, E.K., and Hirst, J.D. 2009. Machine learning in virtual screening. *Comb Chem High Throughput Screen* **12:** 332-343.
64. Meng, X., Yuan, Y., Maestas, A., and Shen, Z. 2004. Recovery from DNA damage-induced G2 arrest requires actin-binding protein filamin-A/actin-binding protein 280. *J Biol Chem* **279:** 6098-6105.
65. Murphy, L.R., Wallqvist, A., and Levy, R.M. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng* **13:** 149-152.
66. Nagano, T., Morikubo, S., and Sato, M. 2004. Filamin A and FILIP (Filamin A-Interacting Protein) regulate cell polarity and motility in neocortical subventricular and intermediate zones during radial migration. *J Neurosci* **24:** 9648-9657.
67. Nakamura, F., Pudas, R., Heikkinen, O., Permi, P., Kilpelainen, I., Munday, A.D., Hartwig, J.H., Stossel, T.P., and Ylanne, J. 2006. The structure of the GPIb-filamin A complex. *Blood* **107:** 1925-1932.

68. Nikki, M., Merilainen, J., and Lehto, V.P. 2002. FAP52 regulates actin organization via binding to filamin. *J Biol Chem* **277:** 11432-11440.
69. Norvell, J.C., and Berg, J.M. 2007. Update on the protein structure initiative. *Structure* **15:** 1519-1522.
70. Ohta, Y., and Hartwig, J.H. 1995. Actin filament cross-linking by chicken gizzard filamin is regulated by phosphorylation in vitro. *Biochemistry* **34:** 6745-6754.
71. Overton, I.M., and Barton, G.J. 2006. A normalised scale for structural genomics target ranking: the OB-Score. *FEBS Lett* **580:** 4005-4009.
72. Overton, I.M., Padovani, G., Girolami, M.A., and Barton, G.J. 2008a. ParCrys: a Parzen window density estimation approach to protein crystallization propensity prediction. *Bioinformatics* **24:** 901-907.
73. Overton, I.M., van Niekerk, C.A., Carter, L.G., Dawson, A., Martin, D.M., Cameron, S., McMahon, S.A., White, M.F., Hunter, W.N., Naismith, J.H., et al. 2008b. TarO: a target optimisation system for structural biology. *Nucleic Acids Res* **36:** W190-196.
74. Ozanne, D.M., Brady, M.E., Cook, S., Gaughan, L., Neal, D.E., and Robson, C.N. 2000. Androgen receptor nuclear translocation is facilitated by the f-actin cross-linking protein filamin. *Mol Endocrinol* **14:** 1618-1626.
75. Passerini, A., Andreini, C., Menchetti, S., Rosato, A., and Frasconi, P. 2007. Predicting zinc binding at the proteome level. *BMC Bioinformatics* **8:** 39.
76. Passerini, A., Punta, M., Ceroni, A., Rost, B., and Frasconi, P. 2006. Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins* **65:** 305-316.
77. Pereira, F., Mitchell, T., and Botvinick, M. 2009. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage* **45:** S199-209.
78. Popowicz, G.M., Schleicher, M., Noegel, A.A., and Holak, T.A. 2006. Filamins: promiscuous organizers of the cytoskeleton. *Trends Biochem Sci* **31:** 411-419.
79. Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., and Sussman, J.L. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21:** 3435-3438.
80. Rivero, F., Koppel, B., Peracino, B., Bozzaro, S., Siegert, F., Weijer, C.J., Schleicher, M., Albrecht, R., and Noegel, A.A. 1996. The role of the cortical cytoskeleton: F-actin crosslinking proteins protect against osmotic stress, ensure cell size, cell shape and motility, and contribute to phagocytosis and development. *J Cell Sci* **109 (Pt 11):** 2679-2691.
81. Robertson, S.P. 2005. Filamin A: phenotypic diversity. *Curr Opin Genet Dev* **15:** 301-307.
82. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science* **229:** 834-838.
83. Rosenberg, S., Stracher, A., and Burridge, K. 1981. Isolation and characterization of a calcium-sensitive alpha-actinin-like protein from human platelet cytoskeletons. *J Biol Chem* **256:** 12986-12991.
84. Salzberg, S. 1995. Locating protein coding regions in human DNA using a decision tree algorithm. *J Comput Biol* **2:** 473-485.

85. Sampson, L.J., Leyland, M.L., and Dart, C. 2003. Direct interaction between the actin-binding protein filamin-A and the inwardly rectifying potassium channel, Kir2.1. *J Biol Chem* **278:** 41988-41997.
86. Sasaki, A., Masuda, Y., Ohta, Y., Ikeda, K., and Watanabe, K. 2001. Filamin associates with Smads and regulates transforming growth factor-beta signaling. *J Biol Chem* **276:** 17871-17877.
87. Seck, T., Baron, R., and Horne, W.C. 2003. Binding of filamin to the C-terminal tail of the calcitonin receptor controls recycling. *J Biol Chem* **278:** 10408-10416.
88. Sells, M.A., Boyd, J.T., and Chernoff, J. 1999. p21-activated kinase 1 (Pak1) regulates cell motility in mammalian fibroblasts. *J Cell Biol* **145:** 837-849.
89. Shifrin, Y., Arora, P.D., Ohta, Y., Calderwood, D.A., and McCulloch, C.A. 2009. The role of FilGAP-filamin A interactions in mechanoprotection. *Mol Biol Cell* **20:** 1269-1279.
90. Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., et al. 2007. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* **35:** D786-793.
91. Sim, K.L., Uchida, T., and Miyano, S. 2001. ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics* **17:** 379-380.
92. Sjekloca, L., Pudas, R., Sjoblom, B., Konarev, P., Carugo, O., Rybin, V., Kiema, T.R., Svergun, D., Ylanne, J., and Djinovic Carugo, K. 2007. Crystal structure of human filamin C domain 23 and small angle scattering model for filamin C 23-24 dimer. *J Mol Biol* **368:** 1011-1023.
93. Smialowski, P., Schmidt, T., Cox, J., Kirschner, A., and Frishman, D. 2006. Will my protein crystallize? A sequence-based predictor. *Proteins* **62:** 343-355.
94. Sodhi, J.S., Bryson, K., McGuffin, L.J., Ward, J.J., Wernisch, L., and Jones, D.T. 2004. Predicting metal-binding site residues in low-resolution structural models. *J Mol Biol* **342:** 307-320.
95. Song, J., and Tang, H. 2004. Accurate classification of homodimeric vs other homooligomeric proteins using a new measure of information discrepancy. *J Chem Inf Comput Sci* **44:** 1324-1327.
96. Stahlhut, M., and van Deurs, B. 2000. Identification of filamin as a novel ligand for caveolin-1: evidence for the organization of caveolin-1-associated membrane domains by the actin cytoskeleton. *Mol Biol Cell* **11:** 325-337.
97. Stossel, T.P., Condeelis, J., Cooley, L., Hartwig, J.H., Noegel, A., Schleicher, M., and Shapiro, S.S. 2001. Filamins as integrators of cell mechanics and signalling. *Nat Rev Mol Cell Biol* **2:** 138-145.
98. Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L., and Geman, D. 2005. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* **21:** 3896-3904.
99. Tigges, U., Koch, B., Wissing, J., Jockusch, B.M., and Ziegler, W.H. 2003. The F-actin cross-linking and focal adhesion protein filamin A is a ligand and in vivo substrate for protein kinase C alpha. *J Biol Chem* **278:** 23561-23569.
100. Tompa, P. 2002. Intrinsically unstructured proteins. *Trends Biochem Sci* **27:** 527-533.
101. Travis, M.A., van der Flier, A., Kammerer, R.A., Mould, A.P., Sonnenberg, A., and Humphries, M.J. 2004. Interaction of filamin A with the integrin beta 7

cytoplasmic domain: role of alternative splicing and phosphorylation. *FEBS Lett* **569:** 185-190.
102. Tseng, Y., An, K.M., Esue, O., and Wirtz, D. 2004. The bimodal role of filamin in controlling the architecture and mechanics of F-actin networks. *J Biol Chem* **279:** 1819-1826.
103. Tu, Y., Wu, S., Shi, X., Chen, K., and Wu, C. 2003. Migfilin and Mig-2 link focal adhesions to filamin and the actin cytoskeleton and function in cell shape modulation. *Cell* **113:** 37-47.
104. Uversky, V.N. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11:** 739-756.
105. Vadlamudi, R.K., Li, F., Adam, L., Nguyen, D., Ohta, Y., Stossel, T.P., and Kumar, R. 2002. Filamin is essential in actin cytoskeletal assembly mediated by p21-activated kinase 1. *Nat Cell Biol* **4:** 681-690.
106. van der Flier, A., and Sonnenberg, A. 2001. Structural and functional aspects of filamins. *Biochim Biophys Acta* **1538:** 99-117.
107. Vucetic, S., Xie, H., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., and Uversky, V.N. 2007. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* **6:** 1899-1916.
108. Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337:** 635-645.
109. Wilson, C.J., Apiyo, D., and Wittung-Stafshede, P. 2004. Role of cofactors in metalloprotein folding. *Q Rev Biophys* **37:** 285-314.
110. Wright, P.E., and Dyson, H.J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293:** 321-331.
111. Wu, C. 2005. Migfilin and its binding partners: from cell biology to human diseases. *J Cell Sci* **118:** 659-664.
112. Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., et al. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34:** D187-191.
113. Xie, Z., Xu, W., Davie, E.W., and Chung, D.W. 1998. Molecular cloning of human ABPL, an actin-binding protein homologue. *Biochem Biophys Res Commun* **251:** 914-919.
114. Yan, C., Terribilini, M., Wu, F., Jernigan, R.L., Dobbs, D., and Honavar, V. 2006. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinformatics* **7:** 262.
115. Yang, Z.R., Thomson, R., McNeil, P., and Esnouf, R.M. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21:** 3369-3376.
116. Yoshida, N., Ogata, T., Tanabe, K., Li, S., Nakazato, M., Kohu, K., Takafuta, T., Shapiro, S., Ohta, Y., Satake, M., et al. 2005. Filamin A-bound PEBP2beta/CBFbeta is retained in the cytoplasm and prevented from functioning as a partner of the Runx1 transcription factor. *Mol Cell Biol* **25:** 1003-1012.

117. Yu, X., Wang, C., and Li, Y. 2006. Classification of protein quaternary structure by functional domain composition. *BMC Bioinformatics* **7:** 187.
118. Yuan, Y., and Shen, Z. 2001. Interaction with BRCA2 suggests a role for filamin-1 (hsFLNa) in DNA damage response. *J Biol Chem* **276:** 48318-48324.
119. Zhang, S.W., Pan, Q., Zhang, H.C., Zhang, Y.L., and Wang, H.Y. 2003. Classification of protein quaternary structure with support vector machine. *Bioinformatics* **19:** 2390-2396.

8. Publications

8.1. Consensus Prediction of Protein Conformational Disorder from Amino Acidic Sequence

Consensus Prediction of Protein Conformational Disorder from Amino Acidic Sequence

Suresh Kumar[1] and Oliviero Carugo*[,1,2]

[1]*Department of Biomolecular Structural Chemistry, Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria*

[2]*Department of General Chemistry, Pavia University, Viale Taramelli 12, I-27100 Pavia, Italy*

Abstract: Predictions of protein conformational disorder are important in structural biology since they can allow the elimination of protein constructs, the three-dimensional structure of which cannot be determined since they are natively unfolded. Here a new procedure is presented that allows one to predict with high accuracy disordered residues on the basis of protein sequences. It makes use of twelve prediction methods and merges their results by using least-squares optimization. A statistical survey of the Protein Data Bank is also reported, in order to know how many residues can be disordered in proteins that were crystallized and the three-dimensional structure of which was determined.

INTRODUCTION

It was recently shown that several proteins do not assume a well defined and stable three-dimensional (3D) structure but are natively unfolded [1]. This was absolutely surprising since unfolded proteins are known to be less stable and soluble *in vitro* and protein misfolding is known to be associated with several conformational diseases, including Parkinson and Alzheimer [2]. However, a considerable fraction of the proteome is constituted by natively unfolded proteins and this fraction seems to be larger in higher organisms than in simpler prokaryotes.

Several techniques to predict conformational disorder in proteins have been designed [3-5] and the performance of many of them is periodically checked, within the CASP initiatives [6], where several blinded predictions are made on targets, the conformational status of which is known only by the CASP organizers and is unknown by the various prediction teams that participate to CASP. In general, it appears that (i) the reliability of these predictions is rather modest and that (ii) different predictions are made by different predictors. The first point is *per se* not surprising, given the intrinsic difficulty of predicting 3D features on the basis of amino acidic sequences. The second point - the inconsistency between different prediction methods - is also not very surprising. In fact, various predictors do not differ only in their algorithms but also in what they define as "conformational disorder" and thus in what they want to predict. For example, in one of the DISEMBL versions [7], all the residues in loops are considered to be conformationally disordered, while in another of the DISEMBL versions, only the residues that were not visible in the crystallographic electron density maps are considered to be disordered. Alternatively, in IUPRED no a priori definition of disorder is used [8]. Despite their limitations, the techniques for predicting conformational disorder are extremely important. Initially, they were designed principally to study the interesting phenomenon of conformation disorder and for large-scale proteome comparisons. Later, it became clear that they have also a series of practical applications, like for example in structural genomics, where they are becoming routine filters in the pipeline of finding suitable targets to be analyzed [3, 4]. In fact, it is obvious that the 3D structure of natively unfolded proteins cannot be determined and that these disordered targets must not be analyzed experimentally by structural biologists.

In this paper we present a consensus method, based on various prediction methods, the performance of which is significantly better than that of each individual predictor. Such a new technique is easily usable with freely available software and is interesting not only for structural genomics initiatives but also for traditional hypothesis-driven structural biology. We also report a statistical survey of the Protein Data Bank that shows the fraction of disordered residues in proteins the crystal structure of which was determined. It appears that a moderate fraction of conformationally disordered residues can be tolerated. About 22% of these crystal structures have more the 5% of the residues that are disordered, though only about 2% of them have more than 20% of the residues in a conformationally disordered status.

METHODS

Data

Information about conformationally disordered proteins was taken from the DISPROT database (http://www. disprot.org/) release 3.3 [9], which lists, in FASTA format, 458 proteins that are known, on the basis of several experimental studies, to be at least partially disordered. Data were downloaded in August 2006. Each residue of these 458 proteins is labeled according to its conformational status: ordered, disordered, unknown. The main advantage of the DISPROT database is that it is curated by experts and it is not based on some automatic procedure. It is thus reasonable to suppose that it contains a very limited number of inaccuracies.

*Address correspondence to this author at the Department of Biomolecular Structural Chemistry, Max F. Perutz Laboratories, Vienna University, Campus Vienna Biocenter 5, A-1030 Vienna, Austria;
E-mail: oliviero.carugo@univie.ac.at

1874-091X/08

Table 1. Individual Prediction Methods Used in the Present Paper

Method	URL	reference
DISEMBL_hot_loops	http://dis.embl.de/	[7]
DISEMBL_loops	http://dis.embl.de/	[7]
DISEMBL_remark465	http://dis.embl.de/	[7]
DISOPRED	http://bioinf.cs.ucl.ac.uk/disopred/	[19]
DRIPRED	http://www.sbc.su.se/~maccallr/disorder/	[20]
FOLDINDEX	http://bip.weizmann.ac.il/fldbin/findex	[21]
GLOBPLOT_B	http://globplot.embl.de/	[22]
GLOBPLOT_r	http://globplot.embl.de/	[22]
IUPRED_L	http://iupred.enzim.hu/	[8]
IUPRED_S	http://iupred.enzim.hu/	[8]
PRELINK	http://genomics.eu.org/spip/PreLink	[23]
RONN	http://www.strubi.ox.ac.uk/RONN	[24]

Individual Predictors

12 individual predictors were used (Table **1**)**.** Some of them are different versions of the same basic algorithm. For example, IUPRED has two versions, one specialized in predicting short disordered polypeptide fragments and the other focused on the prediction of long disordered polypeptide fragments. Others have even three versions, like DISEMBL, which can predict if a residue is in a loop, in a "hot" loop (characterized by high crystallographic B factors), or if it was not observed in the electron density maps, as stated on the "REMARK 465" lines of the files of the Protein Data Bank. Given that the present manuscript is not focused on a particular type of disorder but it is focused on the identification of protein constructs that cannot be studied by structural biologists, we did not make any difference between the various versions of the predictors and we used all of them. This is justified by the fact that we do not want to design a new predictor but we want only to make consensus predictions that can be useful in structural biology for high-throughput structural genomics initiatives and, more in general, in any structural biology project. Moreover, the mathematical approach we used (see below) is essentially unaffected by the use of similar or redundant prediction methods given that it is a least-squares optimization, which by definition, weights all contributions as a function of each other.

Consensus Predictions

Each prediction method (Table **1)** produces binary results: a residue can be predicted to be conformationally ordered or disordered. From a numerical perspective, this can be represented by a value of +1 if it is predicted to be disordered, or by a value of -1, if it is predicted to be ordered. The numerical value of 1 and its sign, positive or negative, are purely arbitrary and different values or opposite signs would not affect the quality of the results.

If one want to use the prediction of several, individual methods and combine their results, it is possible to use least-squares methods to determine the optimal values of the elements x_i of the vector $\mathbf{X}^T=\{x_1, x_2, ..., x_{12}\}$ used in the equation

$$\mathbf{P} \cdot \mathbf{X} = \mathbf{D} \quad (1)$$

where **P** is a N x 12 matrix the elements p_{ij} of which are either +1, if the i^{th} residues is predicted to be disordered by the j^{th} prediction method, or -1 if it is predicted to be ordered, and where **D** is a vector of N elements d_i, the values of which can be either +1, if the i^{th} residues is disordered in the reality, or -1, in the opposite case. The value of N is the total number of residues that are annotated to be ordered or disordered in the DISPROT database and is equal to 54012 residues.

Once the optimal values of the elements of **X** have been determined, it is possible to use them to predict if a residue is conformationally ordered or disordered by computing its *p_cons* value

$$p\_cons = \sum_{i=1}^{12} x_i \cdot p\_indi_i \quad (2)$$

where the values of p_indi_i are either +1, if the residue is predicted to be disordered by the i^{th} prediction method, or -1, if it is predicted to be ordered. If *p_cons* is closer to +1 than to -1, which means if it is greater than 0, the residue is predicted to be disordered. On the contrary, it is predicted to be ordered if *p_cons* < 0. The optimal values of the coefficients x_i are reported in Table **2**.

Prediction Validation

Given the extremely high number (54012) of amino acid residues contained in the DISPROT database, a complete cross-validation, known also as Jack-knife test, is impossible. We performed thus a 20-fold cross-validation: we built randomly 20 non-overlapping sets of residues, each containing 5% of the data, and the optimization of the **X** vector was performed 20 times by discarding each time one of the small subsets, which was then used to compute the *p_cons* values.

Such a separation between the learning sets and the test sets allows one to make unbiased predictions, which can then be compared with the experimentally known conformational statuses of the residues.

Table 2. Optimal Values of the Coefficients x_i to be Used to Compute the *p_cons* Values (Equation 2)

Method	x
DISEMBL_hot_loops	-0.101
DISEMBL_loops	0.377
DISEMBL_remark465	-0.172
DISOPRED	0.048
DRIPRED	0.096
FOLDINDEX	0.262
GLOBPLOT_B	-0.199
GLOBPLOR_r	0.162
IUPRED_L	0.041
IUPRED_S	-0.126
PRELINK	0.078
RONN	0.141

A residue correctly predicted to be disordered was counted as a true positive (tp). A residue correctly predicted to be ordered was counted as a true negative (tn). A disordered residue predicted to be ordered was counted as a false negative (fn). An ordered residue predicted to be disordered was counted as a false positive (fp). Given these four quantities, the prediction reliability was estimated with a series of figures of merit: the sensitivity, the specificity, the accuracy, and the probability excess, defined as

$$sensitivity = \frac{tp}{tp + fn} \quad (3a)$$

$$specificity = \frac{tp}{tp + fp} \quad (3b)$$

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3c)$$

$$probability\_excess = sensitivity + specificity - 1 \quad (3d)$$

The values of these figures of merit can range from 0 to +1 and larger values, closer to +1, are associated with better predictions. It must be observed that some of these figures of merit, typically the accuracy, can be seriously biased if the data are unbalanced. This is exactly what happens here, since the number of ordered residues (2649) is very different from the number of disordered residues (51363) in the database DISPROT. The values of accuracy are thus provided in the present paper only because this figure of merit is used very commonly in computational biology. A much more robust indicator of prediction quality is the probability excess.

RESULTS AND DISCUSSION

Besides their basic biological importance, predictions of protein conformational disorder are important in structural biology, where "impossible" targets must be identified before inserting them in the experimental pipe-line that goes from cloning to structural determination. This is particularly important not only in structural genomics initiatives, the success rate of which is still rather modest, but also in traditional hypothesis-driven applications, especially when the protein construct must be designed by the scientists, like for example in multi-domain protein and viral poly-proteins [10].

Predictions of conformational disorder are thus one of the bioinformatics filters that must be used before moving towards experimental analyses. Other filters are focused on the quaternary structural requirements of a protein chain [11], on protein solubility and stability [12, 13], and some web-based servers were created to assist the users in this task [14, 15].

However, before doing predictions of conformational disorder it is necessary to know what level of disorder can be tolerated by well folded proteins. In fact, while it is clear that the 3D structure of a completely disordered protein cannot be determined, it is also clear that many (or, maybe, most) proteins are partially disordered.

For example, many loops at the protein surface are very flexible and tend to adopt more than a single shape. For this reason, we scanned the Protein Data Bank (PDB) [16, 17] looking for regions conformationally disordered.

This information was extracted from the records labeled with "REMARK 465", where the depositors of the crystal structures declare, if necessary, which residues were not observed in the electron density maps. This analysis was limited to the crystal structures, which are nevertheless the large majority of the entries of the PDB, and it was assumed that the location of completely unfolded segments cannot be detected in the electron density maps.

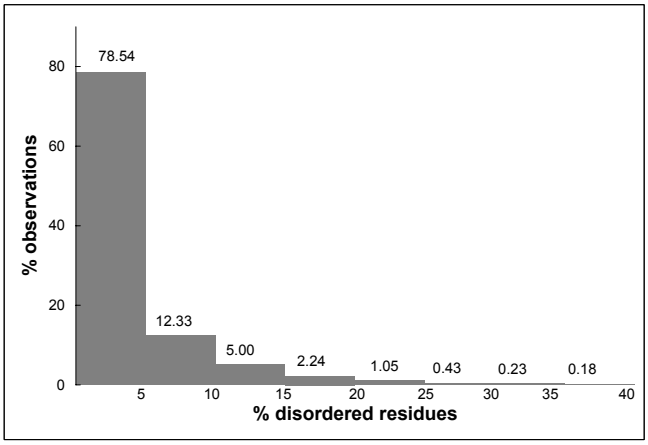


Fig. (1). Distribution of protein crystal structures as a function of the percentage of disordered residues they contain. The data were taken from the Protein Data Bank; a residues was considered to be disordered if not observed in the crystallographic electron density maps; the total number of residues was taken from the SEQRES record of the PDB files.

Fig. (**1**) shows the distribution of the PDB entries according to their fraction of residues not observed which are likely to be conformationally disordered. It appears that a considerable number of structures have conformational disorder. In 22% of them, more than 5% of the residues are disordered.

However, only about 2% of the crystal structures contain more than 20% of the residues that lack a well defined structure. The most extreme case is the entry 1VCR, the light-harvesting complex from *Pisum sativum* thylacoid membrane, where 56% of the residues were not observed, though this crystal structure was determined and refined at very low resolution (9.5 Å) [18].

Fig. (**2**) shows the relationships between the crystallographic resolution and the percentage of disordered residues. It can be seen that resolution tends to decrease if the amount of disorder increases, though the effect of disorder on resolution is not spectacular. In fact the average resolution decreases only from 2.13 to 2.45 Å if the disorder fraction increases from 2.5 to 32.5%.

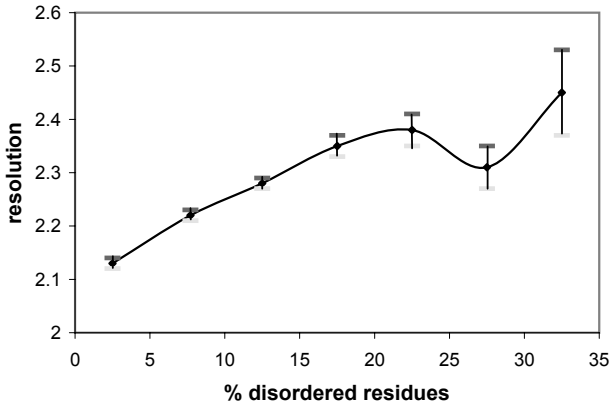


Fig. (2). Dependence between the crystallographic resolution and the percentage of disordered residues observed in the crystal structures deposited in the Protein Data Bank. Vertical bars indicate the standard deviation of the mean.

This clearly shows that protein 3D structures are often partially disordered and that a moderate fraction of conformationally disordered residues can be tolerated. Keeping this in mind, one can now try to predict if a protein has a reasonable probability to be suitable for a structural biology analysis.

We designed a prediction method that is based on several individual prediction algorithms. The only necessary input is the amino acidic sequence of the protein and all the predictors are freely available. Each prediction algorithm must be used separately (Table **1**) and its results must be inserted into equation (2), together with the coefficients x_i reported in Table **2**. If the value of *p_cons* is positive, the residue is predicted to be disordered and if it smaller than zero, the residue is predicted to be ordered. This can easily be done for each residue and, as a consequence, it is possible to reach a global picture of the conformational status of the protein.

This new prediction method, which is essentially a weighted consensus approach, performs quite well, better than any individual prediction algorithm. Table **3** shows the values of several figures of merit, obtained with a 20-fold cross validation procedure. It can be seen that predictions are very accurate, with all the figures of merit larger than 80%. This is impossible by using individual predictors, though all of them have very high specificity. The probability excess, which is the best figure of merit because little influenced by the fact that the data are unbalanced, is equal to 80.1%, a value much larger than any other predictor.

It must be observed that the prediction reliability described above is based on the particular set of proteins available at the DISPROT database. Therefore, it would not be surprising to obtain other estimations of reliability by using different data.

Table 3. Performance of the New Prediction Methods Described in the Present Paper Compared to the Individual Prediction Methods of Table 1

Method	sensitivity	specificity	accuracy	probability excess
Consensus	0.833	0.968	0.814	0.801
DISEMBL_hot_loops	0.481	0.974	0.494	0.455
DISEMBL_loops	0.761	0.966	0.747	0.727
DISEMBL_remar465	0.409	0.977	0.428	0.385
DISOPRED	0.568	0.994	0.586	0.562
DRIPRED	0.640	0.975	0.642	0.615
FOLDINDEX	0.688	0.981	0.691	0.669
GLOBPLOT_B	0.421	0.990	0.445	0.410
GLOBPLOR_r	0.589	0.979	0.597	0.568
IUPRED_L	0.609	0.993	0.624	0.602
IUPRED_S	0.529	0.996	0.550	0.524
PRELINK	0.512	0.970	0.521	0.483
RONN	0.634	0.985	0.642	0.618

As a consequence, the reliability indicators shown in Table **3** cannot be used to rank various prediction methods according to their performances. It is however clear that the consensus approach presented in this manuscript is likely to be superior to all the individual methods on which it is based and it is also reasonable to suppose that an increase of experimental knowledge, which is likely to occur in the future, will allow more accurate predictions.

ACKNOWLEDGEMENTS

This work was supported by the Austrian GEN-AU project BIN-II. Svetlana Kirillova is gratefully acknowledged for her comments.

REFERENCES

[1] Fink, A.L. *Curr. Opin. Struct. Biol.*, **2005**, *15*, 35-41.
[2] Lee, C.; Yu, M.H. *J. Biochem. Mol. Biol.*, **2005**, *38*, 275-280.
[3] Ferron, F.; Longhi S.; Canard, B.; Karlin, D. Proteins., **2006**, *65*(1), 1-14.
[4] Bourhis, J.M.; Canard, B.; Longhi, S. *Curr. Prot. Pept. Sci.*, **2007**, *8*, 135-149.
[5] Dosztanyi, Z.; Sandor, M.; Tompa, P.; Simon, I. *Curr. Prot. Pept. Sci.*, **2007**, *8*, 161-171.
[6] Jin, Y.; Dunbrack, R.L. *Proteins*, **2005**, *61*, Suppl. 7, 167-75.
[7] Linding, R.; Jensen, L.J.; Diella, F.; Bork, P.; Gibson, T.J.; Russell, R.B. *Structure (Camb)*, **2003**,*11*(11), 1453-9.
[8] Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. *Bioinformatics*, **2005**, *21*, 3433-3434.
[9] Vucetic, S.; Obradovic, Z.; Vacic, V.; Radivojac, P.; Peng, K.; Iakoucheva, L.M.; Cortese, M.S.; Lawson, J.D.; Brown, C.J.; Sikes, J.G.; Newton, C.D.; Dunker, A.K. *Bioinformatics*, **2005**, *21*, 137-140.
[10] Carugo, O. *Curr. Protein Pept. Sci.*, **2007**, *8*, 119-120.
[11] Carugo, O. *J. Appl. Cryst.*, in the press.
[12] Smialowski, P.; Martin-Galiano, A.J.; Mikolajka, A.; Girschick, T.; Holak, T.A.; Frishman, D. *Bioinformatics*, **2006**.
[13] Idicula-Thomas, S.; Kulkarni, A.J.; Kulkarni, B.D.; Jayaraman, V.K.; Balaji, P.V. *Bioinformatics*, **2006**, *22*, 278-284.
[14] Smialowski, P.; Schmidt, T.; Cox, J.; Kirschner, A.; Frishman, D. *Proteins*, **2006**, *62*(2), 343-55.
[15] Rodrigues, A.P.; Grant, B.J.; Hubbard, R.E. *Nucleic Acids Res.*, I**2006**, *34(*Web Server issue), W225-30.
[16] Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.; Meyer, E.F.; Brice, Jr., M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J Mol Biol*, **1977**, *112*(3), 535-42.
[17] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. *Nucleic Acids Res.*, **2000**, *28*(1), 235-42.
[18] Hino, T.; Kanamori, E.; Shen, J.R.; Kouyama, *T. Acta Cryst.*, **2004**, D60, 803-809.
[19] Ward, J.J.; Sodhi, J.S.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. *J. Mol. Biol.*, **2004**, *337*, 532-645.
[20] MacCallum, M.R. *Bioinformatics*, **2004**, *20*, i224-i231.
[21] Prilusky, J.; Felder, C.E.; Zeev-Ben-Mordehai, T.; Rydberg, E.H.; Man, O.; Beckmann, J.S.; Silman, I.; Sussman, J.L. *Bioinformatics*, **2005**, *21*, 3435-3438.
[22] Linding, R.; Russell, R.B.; Neduva, V.; Gibson, T.J. *Nucleic Acids Res.*, **2003**, *31*, 3701-3708.
[23] Coeytaux, K.; Poupon, A. *Bioinformatics*, **2005**, *21*, 1891-1900.
[24] Yang, Z.R.; Thomson, R.; McNeil, P.; Esnouf, R.M. *Bioinformatics*, **2005**, *21*, 3369-3376.

Received: November 12, 2007 Revised: November 26, 2007 Accepted: December 12, 2007

8.2. Protein Domain Boundary Predictions: A Structural Biology Perspective

Protein Domain Boundary Predictions: A Structural Biology Perspective

Svetlana Kirillova[a], Suresh Kumar[a] and Oliviero Carugo*[,a,b]

[a]*Department of Biomolecular Structural Chemistry, Max F. Pertuz Laboratories, Vienna University, Campus Vienna, Biocenter 5, A-1030, Vienna*

[b]*Department of General Chemistry, Pavia University, Viale Taramelli 12, I-27100 Pavia, Italy*

Abstract: One of the important fields to apply computational tools for domain boundaries prediction is structural biology. They can be used to design protein constructs that must be expressed in a stable and functional form and must produce diffraction-quality crystals. However, prediction of protein domain boundaries on the basis of amino acid sequences is still very problematical. In present study the performance of several computational approaches are compared. It is observed that the statistical significance of most of the predictions is rather poor. Nevertheless, when the right number of domains is correctly predicted, domain boundaries are predicted within very few residues from their real location. It can be concluded that prediction methods cannot be used yet as routine tools in structural biology, though some of them are rather promising.

INTRODUCTION

Computational/mathematical approaches, such as structural bioinformatics [1], structural class prediction [2, 3], molecular docking [4-9], molecular packing [10, 11], pharmacophore modelling [12], Mote Carlo simulated annealing approach [13], diffusion-controlled reaction simulation [14], graph/diagram approach [15-21], bio-macromolecular internal collective motion simulation [22], QSAR [23-25], protein subcellular location prediction [26-30], protein structural class prediction [31, 32], identification of membrane proteins and their types [33], identification of enzymes and their functional classes [34], identification of proteases and their types [35], protein cleavage site prediction [36-38], and signal peptide prediction [39, 40] can timely provide very useful information and insights for both basic research and drug design and hence are widely welcome by science community.

Several computational approaches aimed to the prediction of protein domain boundaries have been published during the last few years [41, 42]. Besides their intrinsic interest in genome analysis and evolution studies, they are tools that structural biologists may use to optimize the design of the constructs of the proteins, the three-dimensional (3D) structure of which must be determined [43]. While this is particularly important in structural genomics (SG), where the targets have, in general, not been deeply characterized with appropriate biochemical and biophysical tools, this can be important also for traditional hypothesis-driven structural biology projects, where a fine tuning of the construct that is inserted into the experimental pipeline – cloning, expression, purification, etc. – is often necessary in order to get suitable samples [44].

Several information about structure prediction methods are periodically published in the framework of the CASP initiative, the main goal of which is to promote an evaluation of computational prediction methods [45]. This is a periodical exercise, performed every two years since 1994. During CASP experiment a series of protein sequences, the 3D structure of which was determined experimentally though it was not yet published, are distributed to research groups that develop computational methods for predicting protein structural features. It is thus a blinded test, where several methods of "in silico" structural biology techniques can be compared to the reality and to each other. Nevertheless, in each CASP run, the number of targets is obviously quite limited and a prediction method that performs very well in CASP is not necessarily better than other techniques in the reality. It is necessary to make additional investigations focusing on the possibility to use these prediction methods for practical application in structural biology.

Although it is impossible to consider it a rule, it is generally easier to work with single-domain proteins than with multi-domain proteins, since the latter ones tend to be conformationally more flexible [46]. For example, the reciprocal orientation of the domains can vary and depend on the presence of other molecules. Multi-domain proteins may also be little prone to refold if, by chance, they had been overexpressed in cells lacking proper chaperones. This does not mean that multi-domain proteins cannot be studied but it implies that some care must be paid in structural biology experiments and that longer time and larger funding can be expected to be necessary to solve multi-domain proteins. It is thus extremely important to be able to predict, on the basis of its amino acid sequence, if a protein contains one or more structural domains.

CASP is divided into several sections, ranging from prediction of conformational disorder to tertiary structure prediction. Protein domain boundary predictions began to be included in the CASP initiative in 2004. The dissection of a protein into separate structural domains is in fact not trivial at all [46, 47]. It is related to the ill-definition of what a protein domain is. An amino acid segment can be in fact consid-

*Address correspondence to this author at the Department of General Chemistry, Pavia University, Viale Taramelli 12, I-27100 Pavia, Italy; Tel: +43 1 4277 52208; E-mail: oliviero.carugo@univie.ac.at

1874-091X/09

ered to be a structural domain if i) it is a compact ensemble of atoms/residues; ii) it is an ensemble of atoms/residues that behaves as a rigid body, in the sense that it can move relative to other protein moieties without changing its shape; iii) it is a self-folding subunit; iv) it is a polypeptide segment well conserved during molecular evolution. Given the ambiguity in any quantitative definition, the real domain boundaries were defined according to the CASP7 organizers and assessors [47]. They found a reasonable consensus definition for each investigated protein, which seems to be well suitable for a structural biology analysis.

The present study is attempted to compare modern approaches for predicting protein domain boundaries and to define new prediction strategies. Here, we refer to the exercise named CASP7, organized in 2006, for which both predictions and experimental data are available on-line (http://www.predictioncenter.org/casp7/Casp7.html). In this manuscript, several tools, designed for predicting domain boundaries on the basis of the amino acid sequence, will be compared to the real domain architecture. The analysis of these data allows one to answer the following basic questions: i) Is it possible to predict, with the presently available bioinformatics tools, if a protein is made by a single domain or if it contains more than one domain? ii) What is the statistical significance of the available predictions? iii) How accurately can the domain boundaries be predicted in the cases where the presently available bioinformatics predictions work well?

METHODS

Available Data and Tools

Data were obtained from the CASP7 web page (http://predictioncenter.gc.ucdavis.edu/casp7/). Table **1** shows the bioinformatics tools that are freely available and that were used in CASP7. Protein domain prediction methods can be classified into three main categories [42]: i) homology prediction; ii) domain recognition; iii) new domain prediction methods. The 14 prediction methods regarded in present study include all types of approaches. The homology prediction is presented by the chop [48, 49] methods that assign the query sequence to known PDB chains. Dsp [42] uses in addition more general properties of sequence conservation throughout the protein and it can be considered as lying between domain homology and new domain predictions. Domssea [42] belongs to the domain recognition approaches. It is based on the assumption that secondary structure is a more conserved feature of proteins with similar folds than sequence. Domssea aligns the secondary structure predicted for a query protein against a database of 3D domain structures and derives the domain boundaries from the known domain with the most similar secondary structure. Robetta [50] applies BLAST/PSI-BLAST for domain homology prediction and it uses FFAS03 and 3D-Jury to find remote homologues of known domain structure. Hhpred [51] is a server for remote homology detection and for structure prediction using pairwise comparison of profile hidden Markov models (HMMs). In the foldpro [52] method the structural relevance of the query-template pairs is extracted from global profile-profile alignments in combination with predicted secondary structure, relative solvent accessibility, contact map and beta-strand pairing using support vector machines. Distill [53] provides prediction of Contact Density defined as the Principal Eigenvector (PE) of a residue contact map. This information is an important intermediate step towards *ab initio* prediction of protein structure and is used to identify domains. Baker generates 3D protein models using the *de novo* prediction algorithm Rosetta and then assigns domain boundaries using Taylor's structure-based do-

Table 1. Bioinformatics Tools Examined in CASP7 (Names were Taken from CASP)

Tools	URL	Reference
baker	http://robetta.org/submit.jsp	[50]
chop	http://cubic.bioc.columbia.edu/services/chop/index.htm	[48, 49]
chophomo	http://cubic.bioc.columbia.edu/services/chop/index.htm	[48, 49]
distill	http://distill.ucd.ie/distill/	[53]
domfold	http://www.reading.ac.uk/bioinf/DomFold	*
domssea	http://bioinf.cs.ucl.ac.uk/dompred/	[42]
dps	http://bioinf.cs.ucl.ac.uk/dompred/	[42]
foldpro	http://www.igb.uci.edu/servers/psss.html	[52]
hhpred1	http://toolkit.tuebingen.mpg.de/hhpred	[51]
hhpred3	http://toolkit.tuebingen.mpg.de/hhpred	[51]
maopus	http://sigler.bioch.bcm.tmc.edu/CASP7-DOM/	*
metadp	http://meta-dp.cse.buffalo.edu	[54]
NNput	http://webmobis.cs.put.poznan.pl/webmobis/app	*
Robetta	http://robetta.org/submit.jsp	[50]

*- No information provided by authors.

main identification technique. Maopus performs a template screening with PSI-BLAST and FFAS03. The SKELEFOLD approach implemented in Maopus is a *de novo* folding algorithm that uses vector representations of secondary structural elements; domain boundaries are defined with three sequence-based filters. In the domfold method, the output from DomSSEA, DISOPRED and HHsearch is parsed to form a consensus. Metadp [54] and NNput are meta servers that comprise a number of domain prediction methods.

Some of the bioinformatics methods provide multiple predictions. In this case, only the first, which is considered to be the more reliable, was retained for further analysis. Predicted domain boundaries were obtained from the CASP7 web page. The experimental domain boundaries were also obtained from the CASP7 web page, where they were generated by a group of expert scientists. 95 proteins are considered. Given that predictions were not deposited for each protein and for each prediction method, this results in a set of 1210 predictions [47].

Multi-Domain Prediction Using Protein Length

To predict, on the basis of the protein length, that a protein contains one domain or it is a multi-domain construct, a threshold value can be used. If the protein is longer than the threshold value it consists of more than one domain. On the contrary, a protein, smaller than this threshold value, would be predicted to contain only a single domain. Consequently, a true positive (tp) is defined as a multi-domain protein, which is correctly predicted to be a multi-domain protein; a multi-domain protein that is predicted to contain a single domain is defined a false negative (fn); a single-domain protein predicted to be a multi-domain protein is defined a false positive (fp); and a correctly predicted single-domain protein is defined a true negative (tn).

These four types of predictions can be used to estimate the reliability of this prediction methodology. A number of figures of merit have been used for that, like, for example, the Matthews correlation coefficient (mcc) [55]

$$mcc = \frac{(tn \cdot tp) - (fn \cdot fp)}{\sqrt{(fn + tp)(tn + fp)(fp + tp)(fn + tn)}}, \quad (1)$$

the values of which range from -1 to +1 (larger values indicate better predictions) and is little affected by sample heterogeneity (the number of single-domain proteins can be different from the number of multi-domain proteins).

The prediction accuracy was validated with a Jack-knife procedure. In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical applications: independent test dataset, sub-sampling test, and Jack-knife test [56]. However, as elucidated in references [26] and [27], amongst the three cross-validation methods, the Jack-knife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors [57-66].

Statistical Significance of Predictions

To compare the accuracy of different methods with a random prediction we estimated numerically the probability density functions of the indices used to measure the classification validity. This approach is based on idea that the problem of domain boundary prediction using the amino acid sequence is a classification problem. Each residue is in fact predicted to belong to a certain class and it cannot belong to two different clusters at the same time. In other words, a residue can be predicted to belong to a certain domain, to another domain, or to a linker segment. The comparison between a prediction and the reality or between two predictions can thus be performed by using statistical tools that are routinely employed to compare alternative classifications [67] and that are briefly described below.

Given for example two classifications (C and K) of n residues, it is possible to count the number of cases in which residues i and j were classified in the same group in C and K (n_ss), the number of cases in which i and j were classified in the same group in C and in different groups in K (n_sd), the number of cases in which i and j were classified into two different groups in C and in the same group in K (n_ds), and the number of cases in which i and j were classified into two different groups both in C and in K (n_dd). On the basis of this description, it is possible to compute the Jaccard index (J), the Rand coefficient (R), and the Fowlkes-Mallows index (FM), which are defined as:

$$J = \frac{n\_ss}{n\_ss + n\_sd + n\_ds} \quad (2)$$

$$R = \frac{n\_ss + n\_dd}{M} \quad (3)$$

$$FM = \sqrt{\frac{n\_ss}{n\_ss + n\_sd} \cdot \frac{n\_ss}{n\_ss + n\_ds}} \quad (4)$$

where

$$M = n\_ss + n\_sd + n\_ds + n\_dd \cdot \quad (5)$$

By definition, if the two classifications C and K are identical, all the indices (J, R, and FM) are equal to one. It is also important to observe that these indices can be computed independently of the fact that the classifications C and K contain the same number of clusters. This means that the values of J, R, and FM can be computed also if in one case, for example the classification C, all the residues were predicted to be in a unique domain while in the other case, for example the classification K, some residues were assigned to different domains. The only constraint to the computation of J, R, and FM is that both classifications C and K must include the same number of residues, and in the present case this is obvious.

The computation of the values of J, R, and FM is elementary. The estimation of their statistical significance is less obvious [67]. For example, it is difficult to estimate the probability that a certain value of the index J was obtained by chance. From another point of view, if $J_{CK} > J_{DL}$, where J_{CK} monitors the similarity between the classifications C and K and J_{DL} difference between the classifications D and L, it is clear that C and K are more similar to each other than D and L. However, it is more difficult to estimate the statistical significance of the inequality $J_{CK} > J_{DL}$. In other words, it is more difficult to estimate the probability that C and K are

really more similar to each other than D and L. This depends on the fact that the probability density functions of the indices J, R, and FM are unknown and must therefore be estimated numerically on the basis of adequate simulations.

Therefore, we generated a series of simulated partitions, using a Metropolis-Monte Carlo approach, by mean of the following procedure. Each partition is characterized by a series of boundaries that separate a domain and a loop and that can be located also at the N- or at the C-terminus. Given a protein containing N residues, a boundary can be any integer k with $1 \leq k \leq N$. A series of boundaries were generated iteratively. The first (k_0) was randomly selected in the range (1, N); the second (k_1) was randomly selected in the range (1, m_0), where $m_0 = N - k_0$; the third (k_2) was randomly selected in the range (1, m_1) where $m_1 = m_0 - k_1$; and so on, the i^{th} boundary (k_i) was randomly selected in the range (1,m_{i-1}), where $m_{i-1} = m_{i-2}-k_{i-1}$. Two constrains were imposed during the generation of random domain boundaries within a protein. We considered that a domain must contain more than 30 residues and a loop size must be smaller than 30 residues.

10,000 random partitions into domains were generated for proteins containing 75, 100, 125, ..., 550, 575, 600 residues. It was then possible to make 49,995,000 pairwise comparisons between two partitions and the 49,995,000 values of the coefficients J, R, and FM were retained in order to determine their distributions.

As an example, Fig. (**1**) shows the distributions of the index R for some N values. It appears that the distribution dispersion decreases if N increases and that the maximum moves to higher R values for larger proteins. With these data, it is possible to estimate the probability pR to have R values higher than a given value Rx, simply by integrating the probability density curve from Rx to 1, and, analogously, it is possible to get the statistical significance for the other indices.

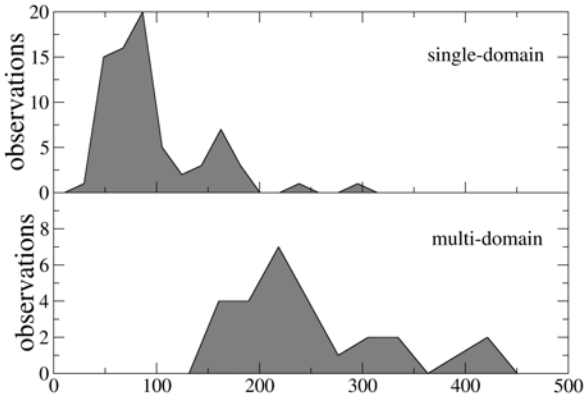


Fig. (1). Distribution of the R index values (fixed bin width of 0.04) computed on 10,000 simulated partitions of proteins containing different number of residues.

Boundary Accuracy

The definition of what is a well predicted domain is obviously arbitrary and here the following conditions were used in order to select the predictions that can be considered to be satisfactory. If the domain contains N residues and it is predicted to contain M residues, and if C is the number of residues that are found in both the real and the predicted domain, a good prediction was defined as a case in which

$$|N - M| < 20 \tag{6}$$

and

$$\frac{C}{\min(M, N)} > 0.95 \tag{7}$$

For well predicted domains, we then computed the differences between the sequence position in which the domain is predicted to begin and the sequence position in which it begins in the reality (Delta_b). Note that a negative value of Delta_b indicates that the domain is predicted to begin before the real beginning along the protein sequence. Analogously, we also computed the differences between the sequence position in which the domain is predicted to end and the sequence position in which it ends in the reality (Delta_e). A positive value of Delta_e indicates that the domain is predicted to be slightly longer, at its C-terminus, than the reality.

RESULTS AND DISCUSSION

Single-Domain Versus Multi-Domain Proteins

Fig. (**2**) shows the distributions of the protein dimensions, measured by the number of amino acid residues, for the single- and multi-domain proteins examined in the CASP7 experiment. As expected, single-domain proteins tend to be smaller than multi-domain proteins, though some overlap between the two distributions exists.

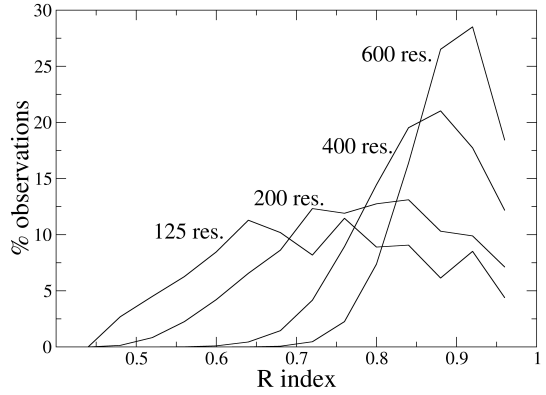


Fig. (2). Distribution of the number of residues (nres) in the single- and muti-domain proteins examined in the CASP7 experiment.

It is thus easy to select a threshold value t and to predict that a protein contains only one domain if smaller than t and that it is multi-domain protein if larger than t. Table **2** shows the mcc values [see equation (1)] observed at various threshold values and validated with a Jack-knife procedure for the proteins examined in the CASP7 experiment. It can be observed that the mcc values are obviously smaller for very small or large values of the threshold. On the contrary they are rather large (>0.6) for intermediate threshold values and

the highest mcc (0.628) is observed with a threshold of 200 residues. This prediction approach is clearly very naive. It simply assumes that a protein domain has a little probability to be very large and, as a consequence, that larger proteins have a higher probability to contain two or more domains.

Table 2. Matthews Correlation (mcc) at Various Threshold Values (t)

t	mcc
70	0.063
80	0.111
90	0.173
100	0.233
110	0.276
120	0.307
130	0.367
140	0.397
150	0.469
160	0.535
170	0.582
180	0.586
190	0.614
200	0.628
210	0.544
220	0.559
230	0.510
240	0.445
250	0.462
260	0.346
270	0.330

A protein is predicted to contain a single domain if it contains less residues that t and it is predicted to contain more than one domain if it has a number of residues larger than t. Data are taken from the proteins examined in the CASP7 experiment.

It is interesting to compare the results of this extremely simple prediction strategy with the results obtained within the CASP7 experiment, where several prediction methods were applied to about 100 proteins. Table **3** shows the mcc values computed on the basis of the predictions deposited by the participants to the CASP7 experiment. The same classification in tp, fp, fn, and tn, which is described in the Methods section, was used. This means that if protein P contains more than a single domain and it was predicted to contain more than a single domain by using the prediction method M, this was considered a true positive (tp). On the contrary, if it was predicted to contain only one domain by the method M, the prediction was considered a false negative (fn), etc. The data of Table **3** clearly show that most of the prediction methods are less reliable than the predictions based on the very simple assumption that a small protein has a high probability to contain a single domain and that a large protein is likely to contain two or more domains. Actually, only four methods (baker, foldpro, maopus and robetta) can predict a multi-domain protein better than the simple predictor (Matthews correlation coefficient larger than 0.628).

Table 3. Matthews's Correlation Coefficients (mcc) Associated with the Prediction of Multi-Domain Proteins by Various Methods Used in the CASP7 Experiment

Method	mcc
baker	0.722
chop	0.178
chophomo	0.230
distill	0.260
domfold	0.262
domssea	0.410
dps	0.277
foldpro	0.840
hhpred1	0.304
hhpred3	0.272
maopus	0.696
metadp	0.189
NNput	0.097
robetta	0.734

What does this mean? Are these bioinformatics tools useless in structural biology? The answer is no. First, some of them seem to be rather accurate. Second, these computational techniques were not specifically trained to identify multi-domain proteins and it is thus not surprising that some of them are not suitable to discriminate mono- and multi-domain proteins. However, it is reasonable to suppose that these bioinformatics tools are still immature and progress should be expected in the future.

Is the Partition Correct?

Table **4** shows the average values of the J, R, and FM indices computed by comparing predicted and real partitions [see equations (2)-(4)]. All the values tend to be large, quite close to their maximal value of 1. However, the probabilities (pJ, pR, and pFM) to observe by chance values higher than these are quite large, ranging from about 30% to about 70%. Baker, foldpro, maopus and robetta are better in predicting a partition that is closer to the real one, with J, R, and FM values that are larger and have a minor probability to be observed by chance. Not surprisingly, they are the same methods that work better to identify multi-domain proteins (see the mcc values of Table **3**).

It must also be observed that matching between prediction and reality is slightly better for small proteins than for large proteins. For example, the probability pJ to find J values larger than those observed by comparing the reality and

Table 4. Average Values of the Indices J,R, and FM and of the Probability pJ, pR, and pFM that a Values Higher than the One that is Observed Might be Obtained by Chance. Standard Deviations of the Mean are Reported in Parentheses

Method	J	R	FM	pJ	pR	pFM
baker	0.80(0.02)	0.82(0.02)	0.88(0.01)	39(4)	35(4)	37(4)
chop	0.66(0.03)	0.70(0.03)	0.79(0.02)	66(5)	63(5)	63(5)
chophomo	0.66(0.03)	0.69(0.03)	0.79(0.02)	67(5)	65(5)	64(5)
distill	0.70(0.02)	0.73(0.02)	0.82(0.01)	58(4)	56(4)	55(4)
domfold	0.76(0.02)	0.77(0.02)	0.86(0.01)	49(5)	48(5)	46(5)
domssea	0.76(0.03)	0.78(0.02)	0.86(0.02)	50(5)	48(5)	48(5)
dps	0.74(0.03)	0.77(0.02)	0.84(0.02)	55(5)	52(5)	52(5)
foldpro	0.82(0.02)	0.84(0.02)	0.90(0.01)	34(4)	32(4)	31(4)
hhpred1	0.77(0.02)	0.78(0.02)	0.86(0.01)	46(4)	45(4)	42(4)
hhpred3	0.76(0.02)	0.78(0.02)	0.86(0.01)	46(4)	45(4)	43(4)
maopus	0.80(0.02)	0.83(0.02)	0.88(0.01)	42(5)	36(5)	39(5)
metadp	0.76(0.03)	0.77(0.03)	0.86(0.02)	49(5)	48(5)	46(5)
NNput	0.71(0.02)	0.73(0.02)	0.83(0.01)	56(4)	55(4)	53(4)
robetta	0.79(0.02)	0.81(0.02)	0.87(0.01)	40(4)	36(4)	37(4)

the predictions of the method "baker" is on average equal to 39%, it decreases to 33% for proteins shorter than 150 residues, and it increases to 43% for proteins containing more then 150 amino acids. This is actually not surprising, since it is easier to predict that a small protein contains a single domain, with, perhaps, two small N- and C-terminal segments protruding from the domain. However, it must be noted that, despite the fact that the pJ, pR, and pFM values can be used only as semi-quantitative indicators - since they are obtained from empirical statistical distributions - it is quite clear that the domain boundary predictions are still quite far from matching the reality.

Are the Domain Boundaries Correct?

We have seen in the previous chapters that the bioinformatics tools are not yet mature enough to be used as routine instruments to design structural biology experiments. However, a very positive feature of these computational methods is that when they work [see equations (6) and (7)] they work very well.

Table 5. Accuracy with which the Domain Boundaries are Identified by Various Prediction Methods

Method	Pc_c	Delta_b	Delta_e
baker	56.2	-1.2(0.3)	2.2(0.5)
chop	26.1	-2.9(1.0)	1.9(0.7)
chophomo	25.0	-2.6(1.0)	2.9(1.0)
distill	33.6	-1.5(0.6)	3.2(0.8)
domfold	38.0	-1.9(0.6)	2.9(0.7)
domssea	42.9	-1.7(0.6)	2.5(0.7)
dps	38.7	-2.2(0.8)	1.6(0.9)
foldpro	62.8	-1.3(0.4)	2.0(0.4)
hhpred1	43.3	-2.1(0.5)	2.6(0.5)
hhpred3	43.4	-2.1(0.5)	2.7(0.5)
maopus	54.2	-1.4(0.6)	3.0(0.8)
metadp	39.8	-1.3(0.7)	3.3(0.7)
NNput	30.8	-1.9(0.7)	2.4(0.8)
robetta	57.9	-1.0(0.3)	1.5(0.5)

The following data are shown: the percentage of domains that are correctly predicted (see text for details) PC_C, the average deviation between the real and the predicted beginning of the domain Delta_b, and the average difference between the real and the predicted end of the domain Delta_e (standard deviations of the mean in parentheses).

Table **5** shows the percentage of domains that are correctly predicted [according to equations (6) and (7)] and the discrepancy between the real and the predicted boundary in the subset of domains that are correctly predicted. It appears that only a relatively modest fraction of the domains can be considered to be well predicted, according to the criteria defined by equations (6) and (7). The percentage of good predictions is about 30-40%, with some prediction methods behaving considerably better than the others and able to well predict about 60% of the domains. The average values of Delta_b (see Methods) are close to and lower than 0 for all the prediction methods. Also the values of Delta_e are very small, though their absolute value tends to be slightly larger than that of Delta_b. Interestingly, the Delta_e values are positive, on average, for each prediction method.

This clearly indicates that in the subset of good predictions the domain boundaries are located with very high accuracy. Actually, a deviation of 1-3 residues is probably a very minor mistake in the process of design a protein construct that has, on average, a high probability to be well folded and conformationally homogeneous. It is also interesting to observe that while the Delta_b mean values are negative, the mean Delta_e values are larger than 0, indicating that predicted domains tend to be slightly longer than real domains.

CONCLUSIONS

In the present manuscript we have analyzed the reliability of the predictions that were made in the CASP7 experiment and that are publicly available. It was found that most of the bioinformatics tools are able to determine if a protein is made by a single domain or if it contains more than one domain, despite a similar reliability is reached by considering only the sequence length, a much simpler strategy. Using a standard and well known statistical test, we showed that most of the predictions that can be done are not impressively better than pseudo-random predictions. It was also observed that although the reliability of the prediction methods seems to be insufficient to make them routine tools in experimental structural biology, their performance can be extremely good. When the domain is correctly identified, its boundaries are very close, within one or two residues, to the experimental ones. In conclusion, these bioinformatics applications are not yet sufficiently accurate to be used as routine tools in experimental structural biology. It is rather probable that the use of more than a single prediction method by a sort of consensus approach might improve the reliability of the predictions. Although these bioinformatics tools are still immature, progress can be expected in the future.

This work was supported by the Austrian GEN-AU project BIN-II. Björn Sjöblom and Kristina Djinovic are acknowledged for helpful discussions. Financial support by Putta None is also acknowledged. One reviewer is acknowledged for a series of references that deserved citation.

REFERENCES

[1] Chou, K.C. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **2004**, *11*, 2105-2134.

[2] Chou, K.C.; Maggiora, G.M. Domain structural class prediction. *Protein Eng.*, **1998**, *11*, 523-538.

[3] Chou, K.C.; Cai, Y.D. Predicting protein structural class by functional domain composition. *Biochem. Biophys. Res. Commun. (Corrigendum*: *ibid*., **2005**, Vol. *329*, 1362), **2004**, *321*, 1007-1009.

[4] Chou, K.C.; Wei, D.Q.; Zhong, W.Z. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem. Biophys. Res. Commun.*, **2003**, *308*, 148-151.

[5] Li, Y.; Wei, D.Q.; Gao, W.N.; Gao, H.; Liu, B.N.; Huang, C.J.; Xu, W.R.; Liu, D.K.; Chen, H.F.; Chou, K.C. Computational approach to drug design for oxazolidinones as antibacterial agents. *Med. Chem.*, **2007**, *3*, 576-582.

[6] Wang, J.F.; Wei, D.Q.; Chen, C.; Li, Y.; Chou, K.C. Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. *Protien Pept. Lett.*, **2008**, *15*, 27-32.

[7] Zhang, R.; Wei, D.Q.; Du, Q.S.; Chou, K.C. Molecular modeling studies of peptide drug candidates against SARS. *Med. Chem.*, **2006**, *2*, 309-314.

[8] Gao, W.N.; Wei, D.Q.; Li, Y.; Gao, H.; Xu, W.R.; Li, A.X.; Chou, K.C. Agaritine and its derivatives are potential inhibitors against HIV proteases. *Med. Chem.*, **2007**, *3*, 221-226.

[9] Zheng, H.; Wei, D.Q.; Zhang, R.; Wang, C.; Wei, H.; Chou, K.C. Screening for new agonists against Alzheimer's disease. *Med. Chem.*, **2007**, *3*, 488-493.

[10] Chou, K.C.; Nemethy, G.; Scheraga, H.A. Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. *J. Am. Chem. Soc.*, **1984**, *106*, 3161-3170.

[11] Chou, K.C.; Maggiora, G.M.; Nemethy, G.; Scheraga, H.A. Energetics of the structure of the four-alpha-helix bundle in proteins. *Proc. Natl. Acad. Sci. USA*, **1988**, *85*, 4295-4299.

[12] Sirois, S.; Wei, D.Q.; Du, Q.S.; Chou, K.C. Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1111-1122.

[13] Chou, K.C. Energy-optimized structure of antifreeze protein and its binding mechanism. *J. Mol. Biol.*, **1992**, *223*, 509-517.

[14] Chou, K.C.; Zhou, G.P. Role of the protein outside active site on the diffusion-controlled reaction of enzyme. *J. Am. Chem. Soc.*, **1982**, *104*, 1409-1413.

[15] Zhou, G.P.; Deng, M.H. An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. *Biochem. J.*, **1984**, *222*, 169-176.

[16] Myers, D.; Palmer, G. Microcomputer tools for steady-state enzyme kinetics. *Bioinformatics*, **1985**, *1*, 105-110.

[17] Chou, K.C. Graphical rules in steady and non-steady enzyme kinetics. *J. Biol. Chem.*, **1989**, *264*, 12074-12079.

[18] Chou, K.C. Rev: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non-steady state systems. *Biophys. Chem.*, **1990**, *35*, 1-24.

[19] Althaus, I.W.; Gonzales, A.J.; Chou, J.J.; Diebel, M.R.; Chou, K.C.; Kezdy, F.J.; Romero, D.L.; Aristoff, P.A.; Tarpley, W.G.; Reusser, F. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. *Can. J. Chem.*, **1993**, *268*, 14875-14880.

[20] Chou, K.C.; Kezdy, F.J.; Reusser, F. Rev: Steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. *Anal. Biochem.*, **1994**, *221*, 217-230.

[21] Andraos, J. Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: New methods based on directed graphs. *Can. J. Chem.*, **2008**, *86*, 342-357.

[22] Chou, K.C. Rev: Low-frequency collective motion in biomacromolecules and its biological functions. *Biophys. Chem.*, **1988**, *30*, 3-48.

[23] Du, Q.S.; Huang, R.B.; Wei, Y.T.; Du, L.Q.; Chou, K.C. Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). *J. Comput. Chem.*, **2008**, *29*, 211-219.

[24] Prado-Prado, F.J.; Gonzalez-Diaz, H.; de la Vega, O.M.; Ubeira, F.M.; Chou, K.C. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *Bioorg. Med. Chem.*, **2008**, *16*, 5871-5880.

[25] Gonzalez-Diaz, H.; Sanchez-Gonzalez, A.; Gonzalez-Diaz, Y. Unified QSAR approach to antimicrobials. Part 3: First multi-tasking QSAR model for Input-Coded prediction, structural back-projection, and complex networks clustering of antiprotozoal compounds. *J. Inorg. Biochem.*, **2006**, *100*, 1290-1297.

[26] Chou, K.C.; Shen, H.B. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.*, **2008**, *3*, 153-162.

[27] Chou, K.C.; Shen, H.B. Rev: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[28] Chou, K.C.; Shen, H.B. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun*., **2006**, *347*, 150-157.
[29] Chou, K.C.; Shen, H.B. Euk-mPLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Protien Res*., **2007**, *6*, 1728-1734.
[30] Chou, K.C.; Shen, H.B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. *J. Protien Res*., **2006**, *5*, 1888-1897.
[31] Chou, K.C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins*, **1995**, *21*, 319-344.
[32] Chou, K.C. Rev: Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci*., **2000**, *1*, 171-208.
[33] Chou, K.C.; Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun*., **2007**, *360*, 339-345.
[34] Shen, H.B.; Chou, K.C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun*., **2007**, *364*, 53-59.
[35] Chou, K.C.; Shen, H.B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information. *Biochem. Biophys. Res. Commun*., **2008**, *376*, 321-325.
[36] Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *J. Biol. Chem*., **1993**, *268*, 16938-16948.
[37] Chou, K.C. Rev: Prediction of HIV protease cleavage sites in proteins. *Anal. Biochem*., **1996**, *233*, 1-14.
[38] Shen, H.B.; Chou, K.C. HIVcleave: A web-server for predicting HIV protease cleavage sites in proteins. *Anal. Biochem*., **2008**, *375*, 388-390.
[39] Chou, K.C.; Shen, H.B. Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem. Biophys. Res. Commun*., **2007**, *357*, 633-640.
[40] Shen, H.B.; Chou, K.C. Signal-3L: A 3-layer approach for predicting signal peptide. *Biochem. Biophys. Res. Commun*., **2007**, *363*, 297-303.
[41] Dovidchenko, N.V.; Lobanov, M.Y.; Galzitskaya, O.V. Prediction of number and position of domain boundaries in multi-domain proteins by use of amino acid sequence alone. *Curr. Protien Pept. Sci*., **2007**, *8*, 189-195.
[42] Bryson, K.; Cozzetto, D.; Jones, D.T. Computer-assisted protein domain boundary prediction using the DomPred server. *Curr. Protien Pept. Sci*., **2007**, *8*, 181-188.
[43] Carugo, O.; Djinovic-Carugo, K.; Gorbalenya, A.E.; Tucker, P. Likelihood of crystallization: Experimental and computational approaches. *J. Appl. Cryst*., **2007**, *40*, 392-393.
[44] Kambach, C. Pipelines, robots, crystals and biology: What use high throughput solving structures of challenging targets? *Curr. Protein Pept. Sci*., **2007**, *8*, 205-217.
[45] Moult, J.; Pedersen, J.T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins*, **1995**, *23*, ii-v.
[46] Carugo, O. Identification of domain in protein crystal structures. *J. Appl. Cryst*., **2007**, *40*, 778-781.
[47] Tress, M.; Cheng, J.; Baldi, P.; Joo, K.; Lee, J.; Seo, J.H.; Baker, D.; Chivian, D.; Kim, D.; Ezkurdia, I. Assessment of predictions submitted for the CASP7 domain prediction category. *Proteins*, **2007**, *69* (Suppl 8), 137-151.
[48] Liu, J.; Rost, B. CHOP proteins into structural domain-like fragments. *Proteins*, **2004**, *55*, 678-688.
[49] Liu, J.; Rost, B. Sequence-based prediction of protein domains. *Nucl. Acids Res*., **2004**, *32*, 3522-3530.
[50] Kim, D.E.; Chivian, D.; Malmstrom, L.; Baker, D. Automated prediction of domain boundaries in CASP6 targets using Ginzu and RosettaDOM. *Proteins*, **2005**, *61*, 193-200.
[51] Söding, J.; Biegert, A.; Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucl. Acids Res*., **2005**, *33*, W244-W248.
[52] Cheng, J.; Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, **2006**, *22*, 1456-1463.
[53] Baú, D.; Martin, A.J.; Mooney, C.; Vullo, A.; Walsh, I.; Pollastri, G. Distill: A suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. *BMC Bioinformatics*, **2006**, *7*, 402.
[54] Saini, H.K.; Fischer, D. Meta-DP: Domain prediction meta-server. *Bioinformatics*, **2005**, *21*, 2917-2920.
[55] Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **1975**, *405*, 442-451.
[56] Chou, K.C.; Zhang, C.T. Rev: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol*., **1995**, *30*, 275-349.
[57] Chen, Y.L.; Li, Q.Z. Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. *J. Theor. Biol*., **2007**, *248*, 377-381.
[58] Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.*, **2007**, *248*, 546-551.
[59] Chen, Y.L.; Li, Q.Z. Prediction of the subcellular location of apoptosis proteins. *J. Theor. Biol*., **2007**, *245*, 775-783.
[60] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol*., **2008**, *252*, 350-356.
[61] Zhang, G.Y.; Fang, B.S. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. *J. Theor. Biol*., **2008**, *253*, 310-315.
[62] Chen, C.; Chen, L.X.; Zou, X.Y.; Cai, P.X. Predicting protein structural class based on multi-features fusion. *J. Theor. Biol*., **2008**, *253*, 388-392.
[63] Du, P.; Li, Y. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. *J. Theor. Biol*., **2008**, *253*, 579-589.
[64] Carugo, O. A structural proteomics filter: Prediction of the quaternary structural type of hetero-oligomeric proteins on the basis of their sequences. *J. Appl. Cryst*., **2007**, *40*, 986-989.
[65] Kumar, S.; Carugo, O. Consensus prediction of protein conformational disorder from amoni acidic sequence. *Open Biochem. J*., **2008**, *2*, 1-5.
[66] Carugo, O. Prediction of polypeptide fragments exposed to the solvent. *In Silico Biol*., **2003**, *3*, 417-428.
[67] Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*, 2nd ed.; Academic Press: San Diego, USA, **2003**.

Received: November 06, 2008 Revised: November 27, 2008 Accepted: November 29, 2008

9. Acknowledgements

I want to express my gratitude to my supervisor and mentor Prof. Kristina Djinovic for her constant support and help and guiding me throughout my initial stages of PhD.

I express my deep sense of gratitude and thanks to my co-supervisor Prof. Oliviero Carugo for his expert guidance, meticulous care, untiring help, encouragement, constant support and consecutive criticism offered during the entire period of my research work. This work would have been rather impossible without his help.

I am extremely thankful to all my lab colleagues for making the lab a cool and friendly place to work.

A special thanks to my friends in Austria and India for their constant support.

I want to express my gratitude to the GEN-AU-BIN-II for providing me with a scholarship.

10. Curriculum vitae

Name	Suresh Kumar Sampathrajan
Date of birth	28.03.1979
Place of birth	Coimbatore, Tamil Nadu
Nationality	Indian
Languages spoken	English, Tamil
Gender	Male
E. mail	sureshbio@gmail.com, suresh.kumar@univie.ac.at
Website	http://www.bioinformaticsweb.org

Educational Qualification

2006-present	Doctoral thesis at the Max F. Perutz laboratories, Department of Structural and Computational Biology, University of Vienna, Dr.Bohrgasse 9, A-1030, Vienna, Austria. Title of dissertation: ‘‘Likelihood of Protein Structure Determination’’ under the supervision of Prof. Dr. Kristina Djinovic, University of Vienna, Austria.
2003-2005	Masters degree, Bharathiar University, Coimbatore,

Tamil Nadu, India (Master of Science).

Title of dissertation: ‘‘Molecular Biodiversity and Banana Genomics: Anonymous markers, Gene Analysis and Access’’ under the supervision of Prof. (Pat) J.S. Heslop-Harrison, University of Leicester, UK.

1998-2002 Bachelor's degree, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu, India (Bachelor of Science).

Suresh Kumar Sampathrajan