# A Meta-learning Approach for Protein Function Prediction

**Dariusz Plewczynski and Subhadip Basu**

**Abstract** One of the major challenges in the post-genomic era is to accurately model the interactions taking place in most cellular processes. Detailed characterization of such interactions is critical for understanding the principles of living cell molecular machinery on the system biology level. This book chapter contains a review of the multiscale protein biological function prediction algorithms that are founded on protein sequence analysis, three-dimensional structure comparison, biological function annotation, and finally molecular interactions. We include diverse computational methods used to predict the biological function for a given biomolecule using multiscale features, and more generally to model a meta-learning prediction system to analyze the impact of micro-dynamics on global behavior for selected biological systems, with important roles in chemistry, biology, and medicine.

## 1 Introduction

Many fundamentally important biological processes are inherently multiscale in nature. These include biophysical phenomena like protein-folding, protein–protein, protein–peptide interactions, protein–ligand docking, and posttranslational modifications (PTMs), to name a few examples. They are intimately coupled to molecular level micro-dynamics, yet their long-time behavior or spatially distant characteristics are related to mesoscale processes. Diverse mathematical methodologies have

D. Plewczynski (✉)
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Pawinskiego 5a Street, 02-106 Warsaw, Poland
e-mail: D.Plewczynski@icm.edu.pl

S. Basu
Department of Computer Science and Engineering, Jadavpur University, Kolkata 32, India
e-mail: subhadip@cse.jdvu.ac.in

P.K. Saha et al. (eds.), *Advanced Computational Approaches to Biomedical Engineering*,
DOI 10.1007/978-3-642-41539-5_5, 

been developed recently for modeling such nonlocal biological processes using varying resolution, or granularity of a model. This book chapter contains a review of the multiscale protein biological function prediction algorithms that are founded on protein sequence analysis, three-dimensional structure comparison, biological function annotation, and finally molecular interactions. The presented approaches span different scales of description of biomolecules, starting from single atoms, ending with a network biology context of biomolecules. We provide a variety of different applications of the proposed methodology to biomedical and biophysical problems. We include diverse computational methods used to predict the biological function for a given biomolecule using multiscale features, and more generally to model a meta-learning prediction system to analyze the impact of micro-dynamics on the global behavior for selected biological systems, with important roles in chemistry, biology, and medicine.

Modern molecular biology provides a vast amount of experimentally verified information, mostly coming from high-throughput studies, such as the Human Genome Project [1, 2], where large-scale sequencing is taking place. In typical drug design procedure pharmaceutical companies perform high-throughput screening studies, where biological activity on the selected protein targets is performed for hundreds of thousands of small chemical molecules. Typical microarray experiments store relative expression profiles for thousands of genes in selected time points. Collecting and verifying currently available experimental data is therefore an important goal of bioinformatics, a rapidly developing computational discipline, which focuses on completing the functional annotation of proteins, inhibitors, DNA/RNA molecules, genes, or more generally all types of biomolecules that can be found in living cells.

Yet, proteins or other molecules are not single, acting-alone entities, they are often working together by forming either stable, permanent complexes with others (like in the case of ribosome), DNA/RNA strains, or ligands (natural metabolites), or interacting transiently with each other, with RNA/DNA molecules, and with small chemical entities. Therefore, a more general term, namely *Interactome*, was introduced that describes the whole universe of molecular interactions in cells, including the protein–protein, protein–ligand, and protein–DNA/RNA interactions. Systems biology addresses those issues, by experimental and theoretical analysis of interactions between biomolecules. Whole proteomes experimental high-throughput techniques rapidly populate available databases with a large amount of average quality data. Such data has to be further validated to qualify each data source contributing to the completion of the interactome. Those collaborative approaches are significantly limited by both time and total cost required for obtaining an accurate and complete map of protein interactions. Therefore, computational methods are beginning to be used to automate the procedure of the careful selection of high quality subsets of available data.

On the other hand, the gap between the number of known protein sequences and the number of crystallized structures is growing rapidly. A three-dimensional structure determines the protein's function; therefore, computational techniques have to be used in order to narrow the gap by predicting structures using only

sequence information. However, despite recent progress in the protein structure prediction community, greatly facilitated by the Critical Assessment of Protein Structure Prediction (CASP) experiment [3], the prediction of the three-dimensional protein structures from their amino acids sequence remains one of the major challenges in modern molecular biology. Therefore, the computational methods have to be designed in order to facilitate the functional annotation of proteins based only on partial biological information describing biomolecules (such as sequences, in the case of proteins). Such *bio-algorithms* are typically tested on a smaller number of examples that are very well characterized by various complementary experimental methods. In the case of proteins, the subset of known proteins with crystallized three-dimensional structures, where both protein sequence and structure is directly linked to performed protein biological function. The ability to learn knowledge on a smaller subset of available data, and apply it on diluted and very noisy biological information, performing data mining during large-scale experimental studies, is the core idea behind machine learning techniques applied in bioinformatics.

Those automatic algorithms nevertheless are focused on characterizing specific, pre-defined features of biomolecules, even if they are trying to study and predict interactions between those components of biological systems. Systems biology is trying to understand how these interactions lead to the function and behavior of that system (for example, the enzymes and metabolites in a metabolic pathway). The living cell is not just the collection of the above-mentioned *life* building blocks, described by their features, or functional annotations. The complex biological systems have to be understood on a higher level, where the organization of those single modules, or biomolecules, is emerges from their individual behaviors, or characteristics. In order to do so, first the interacting agents have to be described, as is done usually in typical low-throughput molecular biology experiments. Secondly, the interactions between those bio-agents have to be characterized, for example, by collecting data in high-throughput studies. Then, computational methods have to be applied in order to combine both levels of description and link agents and their interactions in order to better understand the biological complexity of *life*.

Summarizing, the proteins, genes, small chemical molecules, inhibitors, metabolites are linked in cells with varying scales of molecular interactions, such as protein–protein, protein–ligand, and protein–DNA/RNA. Those interactions are taking place in most cellular processes; therefore, detailed characterization of the interaction repertoire is critical for understanding the principles of living cell molecular machinery on the system biology level. The major challenge in the post-genomic era is to accurately model the organization of those genetic networks, signaling and metabolic pathways, and details of molecular interactions between proteins and their natural or artificial inhibitors, and to understand how they contribute to cellular and organism phenotypes.

### *1.1 The Goal of the Research*

The main goal of the research work presented here is to undertake the challenge of automatically acquiring, storing, organizing, refining, analyzing, and finally building useful working hypotheses for those enormous bioinformatics datasets, especially in the context of multiscale protein function prediction. The integration of those experimental results with previously stored biological knowledge has to be done efficiently, allowing for detection of false or erratic information; both in previously acquired data objects and in newly processed ones. Therefore, new theoretical algorithms have to be introduced, and a new design of the previously used approaches is needed to handle those challenges on the whole genome scale. Due to limitations of traditional machine learning algorithms, various methods of meta-learning or computational intelligence are used to address those bioinformatics problems with very promising results. The importance and successes of those approaches over a diverse range of bioinformatics applications should encourage other scientists to apply these methods to their research.

Computational intelligence methods including traditional machine learning approaches, ensemble methods, artificial neural networks (ANNs), evolutionary algorithms, fuzzy systems, or cognitive computing have been developed during several decades of development. The recent development of CI and the ensemble learning research field is now extensively performed via hybridization and extensions of those algorithms, also in the context of bioinformatics and biomedical applications. This work is aimed at unification of both theoretical algorithms from computer sciences and their applications in biology, with meta-learning based bio-algorithms as the bridge between those two worlds.

## 2 Methods

This work is focused on various interdisciplinary applications of machine learning and data-mining techniques in bioinformatics. The computational methods are used to predict the biological functions for a given biomolecule in various scales, and more generally to model and analyze the selected biological systems. This section includes a description of selected machine learning, clustering, and computational intelligence algorithms, together with their applications in protein function annotation, protein structure prediction, the identification of proteins interactions with small chemical molecules, and finally the analysis of interactions in living cells. The computational methods presented here are based on various machine learning algorithms, dynamic programming methods, several clustering techniques, or similarity searchers. Novel consensus techniques are used to extend classical machine learning algorithms in many applications. Such an approach is inspired by meta-server applications in the protein structure prediction field. Those novel algorithms are now starting to be popular both in biomedical and bioinformatics applications under the general name of meta-approaches or meta-learning.

## 2.1 Meta-learning

The *Meta-learning* term is proposed in the context of computational intelligence as the successful combination of distributed intelligence approaches with traditional machine learning. It is useful in processing of metadata during typical machine learning experiments, when multiple algorithms are used to perform large-scale data mining. The metadata information can include various properties of the learning problem, performance measures of each learning algorithm, the structure of training data, or different patterns that can be deduced from the data. Therefore, the goal is to improve classical learning algorithms by combining different learning algorithms to effectively solve a given learning problem. Such meta-learning approaches are able to better resemble real world problems, allow to significantly improving the overall performance of learning algorithms. The flexibility of each machine learning algorithm is crucial in order to effectively store, organize, and process the acquired data. Each algorithm has its own inductive bias; often it is based on a preselected set of training data features, descriptors, also a set of assumptions about the nature of data. Therefore, it is able to build classification model only if its internal characteristics match the nature of external training data. This means that a learning algorithm may perform very well on one learning problem, but very badly on the next one.

The design of meta-model can be performed at least on three different scales. The microscale meta-learning approach combines several different machine learning algorithms, and builds the consensus between them. In the mesoscale solutions the larger number of independent methods is coupled with different representations of training data by performing feature selection process. In the macro-scale the whole semi-infinite ensemble of learning methods is trained on available data. The construction of meta-learning solution can therefore be drawn in three different layers:

(a) *Microensembles* are constructed using several classification methods combined into single consensus system, for example, by weighted voting procedure [4]. Such method was extensively used by us in applications' papers. We have reported 10 % improvement of error rate over the mean results for the wide range of various interdisciplinary applications.
(b) *Mesoscopic ensembles* may be implemented using standard software, where different machine learning methods are combined with features selection procedure [5]. The decrease of prediction error for the population of learning agents is linked with the distribution of quality of single methods, or statistical influence of selected features on the global meta-result.
(c) *Macroscopic meta-learning* solution can be approximated as semi-infinite learning ensemble, where mean-field theory can be used to get analytical stationary solutions for such system [6, 7].

The applications considered in this work focus on several examples from computational molecular biology that has become increasingly popular in modern research. It provides an excellent overview of machine learning approaches in the context of the complex bioinformatics problems with enormous amount of

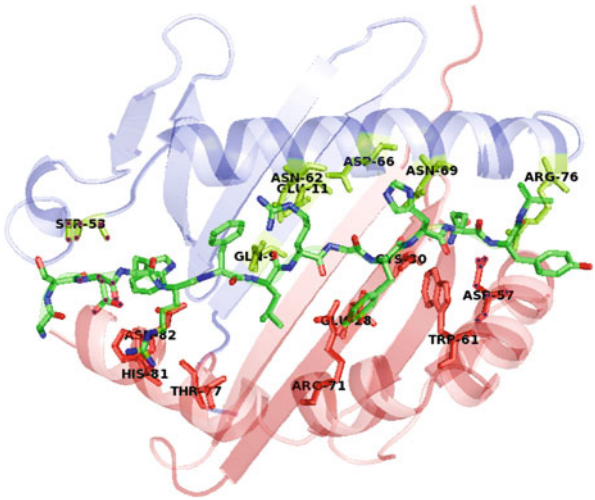


**Fig. 1** Structure of the HLA-DR1 peptide-binding site is illustrated as a sample application of meta-learning prediction system. Top view of the peptide-binding site, with HLA-DR1 residues in contact with the peptide indicated in *yellow* and *red* for the alpha and beta chain, respectively. The peptide residues are colored by atom type, with oxygen in *red*, nitrogen in *blue*, and carbon atoms in *green*. The alpha and beta chains of HLA-DR1 are indicated in *pale-blue* and *pink*, respectively

heterogeneous biological data. Apart from the theoretical foundations of meta-learning, the bioinformatics defines practice of this theoretical framework in the context of many biomedical engineering applications (for an example, see Fig. 1).

One can distinguish three major components of typical bioinformatics workflow, namely the biological data processing, features sampling and preparation, extraction of features (data acquisition), the clustering of relevant features or objects, extraction of similarities between various types of their descriptors (data clustering), and finally the construction of the ensemble of either different machine learning (ML) methods or similar/identical ML methods trained on different subsets of training data, and then the construction of consensus solution that is able to boost theoretical methods quality and precision (meta-learning). The actual treatment of the input training data in the meta-learning procedure is different from previous approaches. The ultimate goal of learning is to discover the relationships between the variables of a system (input, output, and hidden) from direct samples of the system. Most methods assume single representation of training data. Here, during data clustering one builds the set of multiple hypotheses by manipulating the training examples, input data points, target output (the class labels) of the training data, and by introducing randomness into the training data representation. Such approach provides the solid background for more advanced statistical analysis, the background noise extraction, and most informative features selection.

Although quite a few identifiers have been developed in this regard through various approaches, such as clustering, support vector machine (SVM), ANN, or K-nearest neighbor (KNN), and many other classifiers the way they operate the identification is basically individual. Yet, the proper approach usually takes into account the opinions from several experts rather than rely on only one when they are making critical decisions. Likewise, a sophisticated identifier should be trained by several different modes. A consensus of different classifiers often outperforms a single classifier: a learning algorithm searches the hypothesis space to find the best

possible hypothesis. When the size of training set is small, a number of hypotheses may appear to be optimal. An ensemble will average the hypotheses reducing the risk of choosing the wrong one. In addition, most classifiers perform a local search often getting stuck in local optima; multiple starting points provide a better approximation to the unknown function. Finally a single classifier may not be able to represent the true unknown function. A combination of hypotheses, however, may be able to represent this function. This is the core idea of *brainstorming*, which is the core procedure of the Meta-Learning approach. Our consensus approach is similar to other ensemble methods, yet differently from bagging, or boosting, it focuses of the use of heterogeneous set of algorithms in order to capture even remote, weak similarity of the predicted sample to the training cases.

In the following, we present the theoretical framework of the *brainstorming* consensus strategy followed by various applications of different computational intelligence, machine learning, or consensus learning techniques in several practical problems from chemo- and bio-informatics. The foundations of the brainstorming approach, namely the consensus between different types of machine learning algorithms, are described in the context of practical applications to prediction of PTMs of proteins and biological activity of small chemical molecules.

## 2.2 Theoretical Framework of Brainstorming

In general, we define *brainstorming* as a $n-$star quality consensus scheme as $C_n^N$, where $N$ is the number of individual prediction routines or classifiers participating in the specific consensus strategy, and $n$ $(1 \leq n \leq N)$ is the quality of prediction. More specifically, $1-$star prediction says that any one of the possible $N$ classifiers predicts the test sequence to be *positive* for the class type under consideration, and $N-$star represents that *all* classifiers agreed to the decision. Along this principle, we can consider the example of a neural network classifier, and we define the $10-$star quality consensus prediction $C_n^{10}$ as the consensus over ten variations of hidden neurons (neurons in the single hidden layer are varied from 2 to 20 in steps of 2) for a specific performance measure. We designed three different performance measures in our work, one based on the area under the receiver operating characteristic (ROC) curve or the AUC measure, the others being the optimum recall ($R$) and precision ($P$) measures. $C_n^{10}$ is defined over the optimum AUC performance. Similarly, we define $C_n^{20}$ that combines 20 network predictions from AUC and $R$, and $C_n^{30}$ that combines 30 network predictions from AUC, $R$ and $P$ optimized networks. In the following we first discuss the $C_n^{10}$ consensus algorithm and then describe the other variations.

Let $n_k^A$, $n_k^R$, $n_k^P$ be the MLP networks with $K$ neurons in the hidden layer, designed to generate optimum AUC score ($A$), recall ($R$), and precision ($P$) scores, respectively, over the test dataset. Let $p_k^A$, $p_k^R$, $p_k^P$ be the prediction results corresponding to the networks $n_k^A$, $n_k^R$, $n_k^P$ for any unknown test pattern, where:

$$p_k^A = \begin{cases} 1; \text{ test pattern is classified as positive by } n_k^A \\ 0; \text{ otherwise} \end{cases}$$

Similarly, $p_k^R$, $p_k^P$ also generate binary prediction decisions based on the classification confidence of the corresponding MLP classifiers $n_k^R$ and $n_k^P$, respectively. Now the general $n-$star consensus is designed as $C_n^N$, where $n=$ minimum number of networks advocating for a test fragment to be positive. The sum of prediction scores is defined as $S_p^N$. For example, in case of $C_n^{10}$ if $S_p^{10} = \sum_k p_k^{\mathrm{A}}$; $k = 2$ to 20 in steps of 2, a test pattern is said to be predicted with $n-$star quality if $n \le S_p^{10}$. Similarly, for $C_n^{20}$, we estimate $S_p^{20} = \sum_k p_k^A + \sum_k p_k^R$ and for $C_n^{30}$, $S_p^{30} = \sum_k p_k^A + \sum_k p_k^R + \sum_k p_k^P$, where $k=2$ to 20 in steps of 2 in all cases.

In another variation of the brainstorming meta-learning strategy, more consensus prediction models are designed. $C_n^3$ is defined as the consensus among three best $A$, $R$, $P$ networks. $C_n^9$ and $C_n^{12}$ are defined as the consensus over the best networks across different feature sets.

For $C_n^3$ we first define a function Max_AUC_over_Testdata (MAT) to select the best performing network in any given optimization category. The performance is evaluated in terms of maximum AUC score over the unbiased test dataset. Therefore, we first compute $n_{\mathrm{MAT}}^A = \mathrm{MAT}(n_k^A)$; $k=2$ to 20 in steps of 2. Similarly, we compute $n_{\mathrm{MAT}}^R = \mathrm{MAT}(n_k^R)$ and $n_{\mathrm{MAT}}^P = \mathrm{MAT}(n_k^P)$. The corresponding prediction scores are for the three selected networks are defined as $p_{\mathrm{MAT}}^A$, $p_{\mathrm{MAT}}^R$ and $p_{\mathrm{MAT}}^P$, respectively, and the sum of prediction scores as, $S_p^3 = p_{\mathrm{MAT}}^A + p_{\mathrm{MAT}}^R + p_{\mathrm{MAT}}^P$.

In the case of $C_n^9$ we use the MAT function separately for the three different feature sets, viz., HQI-8, HQI-24, and HQI-40 [5, 8]. We define the function $\mathrm{MAT}-\mathrm{HQI}-8$ to generate three best performing nets as $n_{\mathrm{MAT}-\mathrm{HQI}-8}^A = \mathrm{MAT}-\mathrm{HQI}-8(n_k^A)$; $k=2$ to 20 in steps of 2, and likewise $n_{\mathrm{MAT}-\mathrm{HQI}-8}^R$ and $n_{\mathrm{MAT}-\mathrm{HQI}-8}^P$. In the same way three best networks are generated by each of the functions $\mathrm{MAT}-\mathrm{HQI}-24$ and $\mathrm{MAT}-\mathrm{HQI}-40$. The sum of the corresponding prediction scores is then defined as:

$$S_p^9 = p_{\mathrm{MAT}-\mathrm{HQI}-8}^A + p_{\mathrm{MAT}-\mathrm{HQI}-8}^R + p_{\mathrm{MAT}-\mathrm{HQI}-8}^P + p_{\mathrm{MAT}-\mathrm{HQI}-24}^A + p_{\mathrm{MAT}-\mathrm{HQI}-24}^R + p_{\mathrm{MAT}-\mathrm{HQI}-24}^P + p_{\mathrm{MAT}-\mathrm{HQI}-40}^A + p_{\mathrm{MAT}-\mathrm{HQI}-40}^R + p_{\mathrm{MAT}-\mathrm{HQI}-40}^P.$$

Similarly, for $C_n^{12}$ we use four different MAT function separately for the four different feature sets, viz., $\mathrm{MAT}-\mathrm{HQI}-8$, $\mathrm{MAT}-\mathrm{HQI}-24$, $\mathrm{MAT}-\mathrm{HQI}-40$, and $\mathrm{MAT}-\mathrm{AMS3}-10$. The sum of the corresponding prediction scores is then defined as:

$$\begin{aligned} S_p^{12} = {} & p_{\mathrm{MAT}-\mathrm{HQI}-8}^A + p_{\mathrm{MAT}-\mathrm{HQI}-8}^R + p_{\mathrm{MAT}-\mathrm{HQI}-8}^P + p_{\mathrm{MAT}-\mathrm{HQI}-24}^A + p_{\mathrm{MAT}-\mathrm{HQI}-24}^R \\ & + p_{\mathrm{MAT}-\mathrm{HQI}-24}^P + p_{\mathrm{MAT}-\mathrm{HQI}-40}^A + p_{\mathrm{MAT}-\mathrm{HQI}-40}^R + p_{\mathrm{MAT}-\mathrm{HQI}-40}^P \\ & + p_{\mathrm{MAT}-\mathrm{AMS3}-10}^A + p_{\mathrm{MAT}-\mathrm{AMS3}-10}^R + p_{\mathrm{MAT}-\mathrm{AMS3}-10}^P. \end{aligned}$$

As discussed before, $n$-star quality result is obtained for any specific class type between the ANNs in any of the six ways, viz., $C_n^{10}$, $C_n^{20}$ , $C_n^{30}$, $C_n^3$, $C_n^9$, or $C_n^{12}$. We

assign the statistical significance based on "how many ANNs agree that selected fragment is predicted as *Positive* for a given class type."

## 3 Case Studies

### *3.1 Prediction of Proteins Biological Function Using Sequence and Structural Similarity Searches*

The sequence and structural similarity provide the important tool to infer the biological function of a protein. The structural similarity is able to recover twice as more distant relationships than sequence based methods, at the same error rate. Therefore, in [9] work we analyzed the ability of EC (enzyme classification) number prediction that describes the biological function for a given protein [9, 10]. The 3D-Hit structure comparison software developed in [11] provides a unique opportunity to perform fast comparison for a query protein structure in order to find its structural homologs. In another work [12], the Meta-BASIC protein sequence similarity method was modified and applied in order to find sequence homologs for proteins of medicinal relevance. The remote homology detection by Meta-BASIC uses both sequences and predicted secondary structure for similarity assignment. We provided the estimation of the upper limit of the number of molecular targets in the human genome that represent an opportunity for further therapeutic treatment. The druggability is defined here as the ability to bind small-molecule drug, therefore, being likely protein target for screening studies.

The sequence similarity searches were able to find around ∼6,300 human proteins that were similar to sequences of known protein targets as collected from DrugBank database. Therefore, our bioinformatics method estimates the size of druggable human genome to be around 20 % of human proteome. Those predicted protein targets from whole human genome present the unique opportunity for the virtual or experimental high-throughput screening. Figure 2 shows the three-dimensional structure of human glutathione S-transferase in complex with inhibitor.

### *3.2 Consensus Prediction of Protein Secondary Structures Using Two-Stage Multiclass SVMs*

Secondary structure prediction is also crucial for understanding the variety of protein structures and performed biological functions. Prediction of secondary structures for new proteins using their amino acid sequences is of fundamental importance in bioinformatics. We proposed [13] a novel technique to predict protein secondary structures based on position-specific scoring matrices (PSSMs)

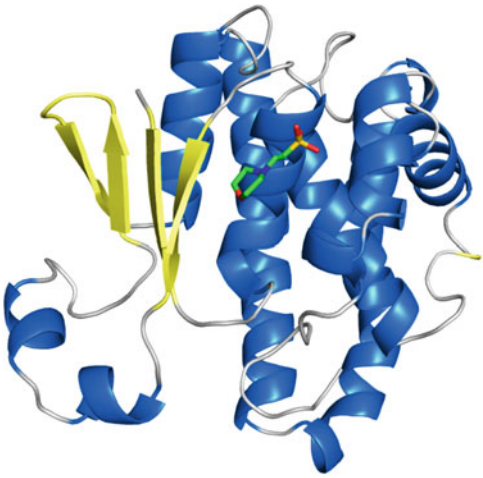

**Fig. 2** Three-dimensional structure of human glutathione S-transferase in complex with inhibitor

and physicochemical properties of amino acids [14]. It is a two-stage approach involving multiclass SVMs as classifiers for three different structural conformations, viz., helix, sheet, and coil. In the first stage, PSSMs obtained from PSI-BLAST [15] and five specially selected physico-chemical properties of amino acids are fed into SVMs as features for sequence-to-structure prediction. Confidence values for forming helix, sheet, and coil that are obtained from the first stage SVM are then used in the second stage SVM for performing structure-to-structure prediction. The two-stage cascaded classifiers (PSP_MCSVM) are trained with proteins from RS126 dataset. The classifiers are finally tested on target proteins of critical assessment of protein structure prediction experiment-9 (CASP9) [3]. PSP_MCSVM with brainstorming consensus procedure performs better than the prediction servers like Predator [16], DSC [17], SIMPA96 [18], for randomly selected proteins from CASP9 targets. The overall performance is found to be comparable with the current state-of-the art tools.

## *3.3 Prediction of PTMs Using Local Sequence Motifs*

We developed a tool that predicts the position of various PTM sites in proteins using only sequence information. Initial study published in [19] was focused on phosphorylation sites identification, and the extended work covering all known types of posttranslational single residue modifications was presented in [20]. Both versions of the method use the ensemble of SVM machine learning algorithms trained using different representations of training objects, i.e. 9-amino acid long fragments of protein sequences dissected around posttranslationally modified sites. Those fragments were extracted from Swiss-Prot database as experimentally confirmed to be phosphorylated by any kinase.

The different representations are build as vectors in a multidimensional abstract space of short sequence fragments linked with different physico-chemical features of amino acids, the sequence profile dissected from the whole protein homology alignments, local structural conformation, etc. Those multiple representations are combined with SVM algorithms (linear, polynomial, and radial kernel functions) to provide statistical models for PTM site predictors.

### *3.4 Automotif Server Version 3.0*

In [4] we replaced the SVM method with ANNs (multilayer perceptron), also simplifying the selection of representations. We presented the recent update of AMS algorithm for identification of PTM sites in proteins based only on sequence information, using ANN method. The query protein sequence is dissected into overlapping short sequence segments. Ten different physico-chemical features describe each amino acid; therefore, nine residues long segment is represented as a point in a 90-dimensional space. The database of sequence segments with experimentally confirmed PTM sites are used for training a set of ANNs. The efficiency of the classification for each type of modification and the prediction power of the method is estimated here using recall (sensitivity), precision values, the area under ROC curves, and leave-one-out cross validation (LOOCV) tests. The significant differences in the performance for differently optimized neural networks are observed, yet the AMS 3.0 tool (http://ams3.bioinfo.pl) integrates those heterogeneous classification schemes into the single consensus scheme, and it is able to boost the precision and recall values independent of a PTM type in comparison with the currently available state-of-the-art methods.

### *3.5 Automotif Server Version 4.0*

The 2011 update of the Auto-Motif Service (AMS 4.0) predicts a wide selection of 88 different types of the single amino acid PTMs in protein sequences [5]. The selection of experimentally confirmed modifications is acquired from the latest UniProt and Phospho.ELM databases for training. The sequence vicinity of each modified residue is represented using amino acids physicochemical features encoded using high quality indices (HQI) obtained by automatic clustering of known indices extracted from AA index database [8]. For each type of the numerical representation, the method builds the ensemble of multilayer perceptron (MLP) pattern classifiers, each optimizing different objectives during the training (for example, the recall, precision, or area under the ROC curve (AUC)). The consensus is built using brainstorming technology, which combines multi-objective instances of machine learning algorithm, and the data fusion of different training objects representations, in order to boost the overall prediction accuracy of conserved short

sequence motifs. The performance of AMS 4.0 is compared with the accuracy of previous versions, which were constructed using single machine learning methods (ANNs, SVM). Our software improves the average AUC score of the earlier version by close to 7 % as calculated on the test datasets of all 88 PTM types. Moreover, for the selected most-difficult sequence motifs types it is able to improve the prediction performance by almost 32 %, when compared with previously used single machine learning methods. Summarizing, the brainstorming consensus meta-learning methodology on the average boosts the AUC score up to around 89 %, averaged over all 88 PTM types. Detailed results for single machine learning methods and the consensus methodology are also provided, together with the comparison to previously published methods and state-of-the-art software tools.

### *3.6 Protein Alignment Method Using Sequence-Structure Motifs*

Defining blocks forming the global protein structure on the basis of local structural regularity is a very fruitful idea, extensively used in description, and prediction of structure from only sequence information. Over many years the secondary structure elements were used as available building blocks with great success. Specially prepared sets of possible structural motifs can be used to describe similarity between very distant, nonhomologous proteins. The reason for utilizing the structural information in the description of proteins is straightforward. Structural comparison is able to detect approximately twice as many distant relationships as sequence comparison at the same error rate.

In our previous paper [21] we provided a new fragment library for local structure segment (LSS) prediction called FRAGlib which is integrated with a previously described segment alignment algorithm SEA. A joined FRAGlib/SEA server provides easy access to both algorithms, allowing a one stop alignment service using a novel approach to protein sequence alignment based on a network matching approach. The FRAGlib used as secondary structure prediction achieves only 73 % accuracy in Q3 measure, but when combined with the SEA alignment, it achieves a significant improvement in pairwise sequence alignment quality, as compared to previous SEA implementation and other public alignment algorithms. The FRAGlib algorithm takes 2 min to search over FRAGlib database for a typical query protein with 500 residues. The SEA service aligns two typical proteins within around 5 min. The joined FRAGlib/SEA server will be a valuable tool both for molecular biologists working on protein sequence analysis and for bioinformaticians developing computational methods of structure prediction and alignment of proteins.

### *3.7 Prediction of Protein–Protein Interaction Prediction Using Domain–Domain Affinities and Frequency Tables*

Protein–protein interactions control most of the biological processes in a living cell. In order to fully understand protein functions, knowledge of protein–protein interactions is necessary. Prediction of PPI is challenging, especially when the three-dimensional structure of interacting partners is not known. We proposed a novel knowledge-based prediction method [22], which predicts interactions between two protein sequences by exploiting their domain information. We trained a two-class SVM on the benchmarking set of pairs of interacting proteins extracted from the Database of Interacting Proteins [23]. The method considers all possible combinations of constituent domains between two protein sequences, unlike most of the existing approaches. Moreover, it deals with both single-domain proteins and multi-domain proteins; therefore, it can be applied to the whole proteome in high-throughput studies. Our machine learning classifier, following a brainstorming consensus approach, achieves accuracy of 86 %, with specificity of 95 %, and sensitivity of 75 %, which are better results than most previous methods that sacrifice recall values in order to boost the overall precision. Our method has on average better sensitivity combined with good selectivity on the benchmarking dataset.

## 4 Summary and Future Directions

First, we performed an extensive review of different physico-chemical features of proteins that have a dramatic impact on their role in a cell. In addition, several computational methods are used for extending the acquired experimental knowledge to unknown or newly sequenced proteins, where no structural information is available. The work sheds some light on details of the fundamental link between a protein's sequence, structure, and its function performed in the living cell. This systems biology approach is further validated on known high quality experimental examples.

In the second step of our analysis, we limited the discussion to those cases where the three-dimensional structures of both a protein and its metabolites, or interaction partners, are known. We evaluated the various algorithms that are trying to predict the correct structure of the protein–protein or protein–ligand complexes, and estimate their binding affinity value by scoring the strength of interactions. The results were later used as the foundation of the effective protocol for docking, within the optimization technique based on multiple linear regression (MLR).

The ultimate goal of those computational approaches is to provide the methodology for automatic characterization of interaction partners, using either amino acids sequences (or full atom representations) and/or three-dimensional structures (if some structural information is known). We focused here on practical applications

of the theoretical methods, also elucidating the software or web interfaces needed to run them in the high-throughput predictions on the whole genome scale. The proposed general validation of theoretical methods on real life, experimental data provides the best estimation of their accuracy. We compared here various clustering, machine learning or statistical methods for bioinformatics knowledge acquisition, processing and mining. In addition, we presented a novel computational intelligence algorithm, namely the brainstorming meta-learning technique applied to various problems from bioinformatics and chemoinformatics. It covered the analysis of interactions of proteins, functional links between protein function, structure and sequence, and other applications in the context of the life sciences.

In this book chapter, we are more focused on applications than on the theoretical foundation of meta-learning. This extensive review of applications of meta-learning is focused on bioinformatics, an enormously rich application field for mathematical methods. The complexity of scientific problems and the large amount of heterogeneous biological data provide an excellent test-ground for machine learning approaches in a real-life context. In return, bioinformatics, while using different theoretical methods, can also offer serious advances in theoretical computational intelligence. Most computational approaches are based on comparative molecular similarity analysis of proteins with known and unknown characteristics.

Eventually, the broader goal of this project is to develop a multiscale computational model of the entire life cycle of living organisms. In this process, both the bottom-up and the top-down approaches need to converge to delineate the underlying functionalities of cellular processes. The real-time analysis of incoming time-dependent non-equilibrium data performed by modern large-scale data-mining techniques, or computational intelligence algorithms, provides the theoretical framework, which in turn allows for better understanding the flexibility of learning mechanisms observed in real biological systems. Recent advances in theoretical neuroscience allow for better understanding of the brain structure, dynamics and performed basic cognitive functions on the biology level (see Fig. 3). We hypothesize that those two presently distant areas of science, namely cognitive science and computational intelligence, will eventually merge into a single research area, where complex and time-dependent meta-learning systems, inspired by the real mammalian brain structure, can be used as the implementation of cognitive systems.

The crucial point here is that in addition to learning algorithms the coupling to sensors has to be provided in order to allow the artificial system (or modeled biological one) to perform the effective cognitive process when discovering the world that surrounds it. The ultimate goal of artificial intelligence studies, i.e., a constantly evolving meta-learner that in real time accumulates acquired information in the form of processed knowledge, is still long way from the present state of research. Both theoretical algorithms and hardware resources (computers or specialized accelerators) have to be improved in order to perform instant, rapid learning using different algorithms, when new input is presented to the system. Only then will the "intelligent" system be able to answer most of our expectations focusing on computational intelligence.

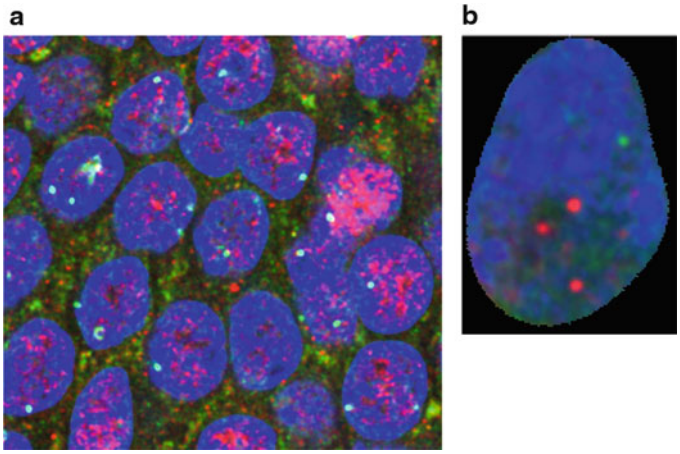


**Fig. 3** (**a**) The confocal images of nuclei in living cells with selected active genes marked by fluorescent proteins. (**b**) Segmented view of a single nucleus. This type of data modality links the molecular view described in our chapter with imaging techniques, applied to living systems. *Data courtesy*: Nencki Institute of Experimental Biology, Warsaw, Poland

## References

1. Watson, J.D.: The human genome project: past, present, and future. Science **248**, 44–49 (1990)
2. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F.: Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252**, 1651–1656 (1991)
3. Moult, J., Fidelis, K., Kryshtafovych, A., Tramontano, A.: Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins **79**, 1–5 (2011)
4. Basu, S., Plewczynski, D.: AMS 3.0: prediction of post-translational modifications. BMC Bioinformatics **11**, 210 (2010)
5. Plewczynski, D., Basu, S., Saha, I.: AMS 4.0: consensus prediction of post-translational modifications in protein sequences. Amino Acids **43**(2), 573–582 (2012)
6. Plewczynski, D.: Mean-field theory of meta-learning. J. Stat. Mech. **11**, P11003 (2009)
7. Plewczynski, D.: Landau theory of meta-learning. In: Security and Intelligent Information Systems, vol. 7053, pp. 142–153. Springer, Heidelberg (2012)
8. Saha, I., Maulik, U., Bandyopadhyay, S., Plewczynski, D.: Fuzzy clustering of physicochemical and biochemical properties of amino acids. Amino Acids **43**, 583–594 (2012)
9. von Grotthuss, M., Plewczynski, D., Ginalski, K., Rychlewski, L., Shakhnovich, E.: PDB-UF: database of predicted enzymatic functions for unannotated protein structures from structural genomics. BMC Bioinformatics **7**, 53 (2006)
10. von Grotthuss, M., Plewczynski, D., Vriend, G., Rychlewski, L.: 3D-Fun: predicting enzyme function from structure. Nucleic Acids Res. **36**, W303–W307 (2008)
11. Plewczyński, D., Paś, J., von Grotthuss, M., Rychlewski, L.: 3D-Hit: fast structural comparison of proteins. Appl. Bioinformatics **1**, 223 (2002)
12. Plewczynski, D., Rychlewski, L.: Meta-basic estimates the size of druggable human genome. J. Mol. Model. **15**, 695–699 (2009)
13. Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., Plewczynski, D.: PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. J. Mol. Model. **17**, 2191–2201 (2011)

14. Kawashima, S., Kanehisa, M.: AAindex: amino acid index database. Nucleic Acids Res. **28**(374) (2000)
15. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**, 3389–3402 (1997)
16. Frishman, D., Argos, P.: Seventy-five percent accuracy in protein secondary structure prediction. Proteins **27**, 329–335 (1997)
17. King, R.D., Sternberg, M.J.E.: Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci. **5**, 2298–2310 (1996)
18. Levin, J.M.: Exploring the limits of nearest neighbour secondary structure prediction. Protein Eng. **10**, 771–776 (1997)
19. Plewczynski, D., Tkacz, A., Wyrwicz, L.S., Rychlewski, L.: AutoMotif server: prediction of single residue post-translational modifications in proteins. Bioinformatics **21**, 2525–2527 (2005)
20. Plewczynski, D., Tkacz, A., Wyrwicz, L.S., Rychlewski, L., Ginalski, K.: AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. J. Mol. Model. **14**, 69–76 (2008)
21. Plewczynski, D., Rychlewski, L., Ye, Y., Jaroszewski, L., Godzik, A.: Integrated web service for improving alignment quality based on segments comparison. BMC Bioinformatics **5**, 98 (2004)
22. Chatterjee, P., Basu, S., Kundu, M.M., Nasipuri, M., Plewczynski, D.: PPI_SVM: prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. Cell. Mol. Biol. Lett. **16**, 264–278 (2011)
23. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.-M., Eisenberg, D.: DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. **30**, 303–305 (2002)