FEBS Open Bio 4 (2014) 310–314

FEBS openbio

journal homepage: www.elsevier.com/locate/febsopenbio

A word of caution about biological inference – Revisiting cysteine covalent state predictions

Éva Tüdős [a], Bálint Mészáros [a], András Fiser [b,c], István Simon [a,*]

[a] Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, P.O. Box 286, H-1519 Budapest, Hungary
[b] Department of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA
[c] Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA

ARTICLE INFO

Article history:
Received 25 November 2013
Revised 5 March 2014
Accepted 7 March 2014

Keywords:
Protein prediction
Cysteine redox state
Protein structure
Prediction accuracies
Biological inference

ABSTRACT

The success of methods for predicting the redox state of cysteine residues from the sequence environment seemed to validate the basic assumption that this state is mainly determined locally. However, the accuracy of predictions on randomized sequences or of non-cysteine residues remained high, suggesting that these predictions rather capture global features of proteins such as subcellular localization, which depends on composition. This illustrates that even high prediction accuracy is insufficient to validate implicit assumptions about a biological phenomenon. Correctly identifying the relevant underlying biochemical reasons for the success of a method is essential to gain proper biological insights and develop more accurate and novel bioinformatics tools.

1. Introduction

The benefits of protein sequence based computational prediction methods are usually twofold. On one hand, they offer a fast and inexpensive way to obtain new biological information about a protein, which then can be used to design follow-up experiments or to explain existing experimental observations. On the other hand, the success of these prediction algorithms is generally considered to validate the underlying hypothesis about principles governing structure formation or biochemical role of the feature being predicted. For instance, the success of early secondary structure prediction methods in the sixties and seventies, based on α-helix and β-strand forming preferences of individual residues indicated that most of the information about the preferred secondary structure of any segment of the protein is encoded in the local sequence itself [1,2]. Although the development of secondary structure predictions encompassed many years, even the most advanced secondary structure prediction algorithms, reaching as high as 80% accuracy, are able to do so by using local sequence information only [3,4]. In this case the basic assumption (the dominance of local effects in secondary structure formation) is in fact in agreement with both bioinformatics and experimental results. The idea itself predated bioinformatics analysis of three dimensional structures and was later validated by the appropriate calculations, confirming that long range interactions indeed played only a secondary role [5,6]. Experimental investigations also demonstrated that isolated segments have a tendency to prefer conformations similar to the one in folded structures [7].

However, in case of other bioinformatics methods the relationship between a successful prediction and the underlying hypothesis may be less trivial. Highly specific residue-level structural or biochemical features are predicted from local sequence patterns; however it is usually not elaborated why certain features are supposed to be encoded in the immediate sequence environment of residues. Here, as a case study, we focus on the development of a prediction method capturing the covalent state of cysteine (Cys) residues from the amino acid sequence. Despite being a highly specific area of bioinformatics research, it can serve with general conclusions about biological inference and the potential pitfalls of interpreting the success of a prediction method as a verification of underlying assumptions.

As protein sequences are obtained primarily by genome sequencing, the location of post-translational modifications – such as disulfide bonds – is usually unknown for the majority of

Abbreviations: Acc, prediction accuracies; FN, false negatives; FP, false positives; PDB, Protein Data Bank; TN, true negatives; TP, true positives

* Corresponding author. Tel.: +36 1 3826295; fax: +36 1 3826710.

E-mail addresses: tudos.eva@ttk.mta.hu (É. Tüdős), meszaros.balint@ttk.mta.hu (B. Mészáros), andras.fiser@einstein.yu.edu (A. Fiser), simon.istvan@ttk.mta.hu (I. Simon).

http://dx.doi.org/10.1016/j.fob.2014.03.003

proteins. Protein structure determination would clearly benefit from the knowledge of the oxidation state of various Cys residues, e.g. if a certain Cys does or does not form a disulfide bond [8–11]. Furthermore, disulfide bond connectivity patterns can be used to discriminate between protein folds and can facilitate the accurate superimposition of protein structures [10]. The underlying assumption in these methods is that similar disulfide bond connectivity patterns place similar spatial constraints on proteins, resulting in similar protein structures. Meanwhile, variation in disulfide bridge patterns within the same superfamily may be used to infer variation of protein function. Information on disulfide formation can be incorporated to enhance molecular simulations of folding [12]. Due to all these critical roles that disulfide bonds play in proteins, there is a long standing interest to predict Cys residues that can form disulfide bonds, or more generally, predict which Cys residues are oxidized and bound to another Cys residue (or to other factors) and which Cys are reduced and have a free, highly reactive thiol group.

During the 1980's the sequence databases had become large enough to enable the recognition of the non-random pairing of residues within their sequential vicinity [13,14]. Early prediction methods were based on the observed different sequence environments of cysteines and half-cystines of the disulfide bonds [15]. The success of a neural network based approach considering the sequence environments of cysteines and half-cystines [16] also suggested that most of the information on the preferred oxidation state of Cys residues must be encoded in the surrounding sequence segments. This hypothesis formed the basis of a prediction approach, where a disulfide bond forming statistical potential was calculated for Cys residues from the relative frequency of residues in the surrounding decapeptides [15]. This method – performing with 71% accuracy – was developed nearly two decades ago and according to citation data of the paper it is still in use; although since then more accurate, albeit more complicated methods have been developed that use a wide range of other input data in addition to the protein sequence [11] (reaching 82% accuracy, e.g. [17]). Furthermore, the incorporation of mass spectrometry data into S–S bond determination methods (see [18–20]) has also significantly improved their accuracy, even surpassing the efficiency of sequence based methods, albeit on smaller datasets [21]. The primary source of information for this study was the Protein Data Bank (PDB [22]), which is approximately 40 times larger now than it was at the time. Hence, it seems plausible that the re-parametrization of the method would significantly increase the accuracy.

By now it is clear that due to the high glutathione concentration in the cytoplasm (due to the constituently active glutathione reductase enzyme), and the oxidative environment in the extracellular matrix, almost all Cys residues in the cytoplasm (intracellular Cys residues) are in the free, thiol form, while almost all Cys residues outside the cytoplasm (in the extracellular matrix or in the various compartments of the cell, extracellular Cys residues) are either in disulfide bonds or in a liganded form [23]. However, it has also become known that the amino acid compositions of intracellular and extracellular proteins are significantly different [17,24–26]. Therefore one can present an alternative hypothesis, according to which the 20 residues in the flanking decapeptides of Cys residues might only identify the extracellular or intracellular nature of the entire protein and the oxidation state of the Cys residues is simply the consequence of the subcellular localization. This idea is supported by the fact that the inclusion of subcellular localization in the prediction algorithms improve their accuracy only by a few points [27] indicating that this information is already largely encoded in other input data (e.g. the sequence).

It is rather common that prediction methods are used for a different purpose than what they were developed for. For example, the above mentioned Cys prediction method [15] was successfully used to design a site directed mutagenesis experiment, to predict where a Cys residue needs to be inserted into a sequence to have a larger probability of forming a disulfide-bond [28] or to find out which of the designed Cys residues form a disulfide bond with a sequentially distant Cys residue [29]. However, in theory, for this type of predictions the method can only be meaningful if there is a significant direct influence of the local sequence environment on the disulfide forming potential.

In this study we explore the origin of the potential of a Cys to exist in a disulfide bond/liganded (bound) form and in thiol (free) form, as calculated from the signal of characteristic local sequence patterns in the neighborhood of Cys residues. Based on these results, we aim to assess the correctness of considering the success of the developed method as a proof of the idea that Cys oxidation state is directly encoded in the local sequence. In order to do this, the original prediction method was re-implemented in this study using the current, 40 fold larger PDB database. We also revisited earlier reports that claim that certain relative sequential positions around Cys residues play specific roles in determining its oxidation state. Subsequently, the method implementation was repeated using shuffled protein sequences to remove all local sequence information. Finally, the specificity of the prediction algorithm was further challenged by implementing it for all the 20 residue types, not only for Cys. The results can either reinforce the view that Cys redox state is mainly determined by local effects or can show the non-causal relationship between this underlying view and prediction accuracy, serving as a general warning about biological inferences concerning bioinformatics methods.

2. Datasets and methods

2.1. Real protein sequence dataset

Entries in Protein Data Bank (PDB [22]) were filtered at 25% sequence identity level using BlastClust (http://www.ncbi.nlm.nih.gov/blast/Blast/) to remove redundant sequences. The resulting dataset contains 3881 polypeptide chains with 19202 Cys residues. Out of these sequences 569, 2039 and 132 contain half-cystines, free cysteines and both half-cystines and free cysteins, respectively. Using a more general classification we also discriminate between oxidized Cys that are bound to either another Cys residue in a disulfide bond or to ligands and free Cys that are not bound. In this classification there are 693 proteins containing bound Cys only, 1804 of them contain only free cysteins and 243 contain both. Our dataset also contains 1117 polypeptide chains that do not contain any Cys residues at all and 24 chains were discarded as the oxidation state of Cys residues could not be determined. According to the UniProt annotations [30] (http://www.uniprot.org/) the 3881 polypeptide chains can be split into three groups: 1783 intracellular, 1024 extracellular and 910 transmembrane proteins. There are 9515 half-cystines and 9687 free cysteines or – using the alternative definition – there are 10996 Cys residues in some sort of bound state (oxidized) and 8206 in free thiol (reduced) state.

2.2. Random protein sequence dataset

Randomly mixed sequences were obtained by shuffling the order of residues within each protein from the *Real protein sequence dataset* individually.

2.3. Disulfide bond forming potentials

First position specific residue preference matrices were calculated. For each position in a ±10 residue window centered on Cys residues the occurring amino acids were counted. Next, the occurrences of all amino acids in these positions were compared between bound and free cysteines. These were expressed in ratios: values greater than 1.0 indicate that the corresponding residue in the given position appears more often in the vicinity of half cystines than in that of free cysteines. After this, potentials were calculated from the position specific residue preference matrices (Supplementary Table 1) according to the method described in [15]. These potentials are obtained as the products of the position specific relative abundances of amino acids around the cysteine residues.

2.4. Cys oxidation prediction

The oxidation states of cysteines were predicted according to SecretomeP2.0 [31] (http://www.cbs.dtu.dk/services/SecretomeP/) predictions. Cysteins of proteins predicted as extracellular were classified as oxidized while cysteins in the intracellular proteins were regarded as reduced.

2.5. Prediction evaluation

Prediction accuracies (Acc) were obtained as a fraction of correctly predicted cases, using the standard formula: Acc = [TP + TN]/[TP + TN + FP + FN], where TP and TN is the number of true positives (correctly identified bound Cys) and true negatives (correctly identified free Cys) and FP and FN are false positives (free Cys incorrectly predicted to be bound) and false negatives, respectively (bound Cys incorrectly predicted to be free).

3. Results

As a first step, the disulfide bond forming potentials described in [15] were recalculated using proteins from the current PDB (see Datasets and methods). Despite the much larger amount of input data, the accuracy of disulfide-bound prediction has not improved notably, the accuracy increased slightly from 71% to 75% (Supplementary Table 2). Testing prediction methods on different datasets presents an inherent difficulty, (as discussed in detail for example for signal peptide predictions here [32]) and therefore this increase in accuracy cannot be deemed significant. We also repeated the calculations to predict the bound-state of cysteine residues and calculated the accuracy of the prediction, with similar results (71% using a jack-knife method).

Once we re-established the method, we revisited the earlier hypothesis, according to which some positions are more important than others in defining the covalent state of Cys residues [15]. We compared the position specific values of the frequency ratios used to calculate the disulfide-bond forming potential. When comparing the pairwise correlation coefficients of the position specific potential vectors (values of the columns of the frequency tables in Supplementary Table 1), it seems that the third sequential neighbors of Cys residues may have some unique preference as opposed to all other positions (Supplementary Table 3). This effect can be attributed to the role of the ±3rd positions in coordinating metal-ion binding (Fig. 1). To test this hypothesis, we repeated the disulfide bond forming prediction method taking into account the frequency of occurrences in the ±3rd neighboring amino acid positions only and achieved a 59% accuracy (we obtained a similar 59% accuracy when tested for bound state prediction). Sequentially closer residues usually have more non-random distribution [14], probably due to certain structural constraints, therefore it is expected that taking into account only the nearest residues for the development of the prediction method may achieve a higher prediction accuracy than using only more distant ones. To explore this, all the possible sequential ranges were considered in subsequent calculations from 1 to 10 neighboring positions and the power of predictions were tested (Supplementary Table 4). Although accuracies increase with the widening of the considered sequence window, a "diminishing returns" effect can be seen. Accuracies seem to reach a plateau suggesting that more distant positions contribute less and less information.

Next, we directly challenged the core idea that local sequence information is indeed the source of successful prediction. We performed the training of the prediction method on randomized sequences. In this case the specific information about the local sequence environment of Cys residues is supposedly completely removed, however the overall sequence compositions of the test proteins are retained. Surprisingly, the prediction accuracies on these randomly shuffled sequences remained just as good as on native sequences: 71% and 69% for the disulfide bridge forming and oxidized Cys, respectively (Supplementary Table 2).

Another way to directly test the hypothesis that the oxidation state of Cys residues is determined by the local sequence is to repeat the development of the prediction method, but this time applying it to the rest of the 19 amino acids besides Cys (Supplementary Table 5). We calculated the "disulfide-bond forming" preferences (and the bound-state preferences) for all the 19 amino acids at the locations of the cysteines of the protein studied. In this case the "bound" state definition of each residue was applied through the context of co-occurrence of these residues and oxidized Cys in the same protein. For instance, if a residue was sampled from a protein that had only oxidized Cys, then any of

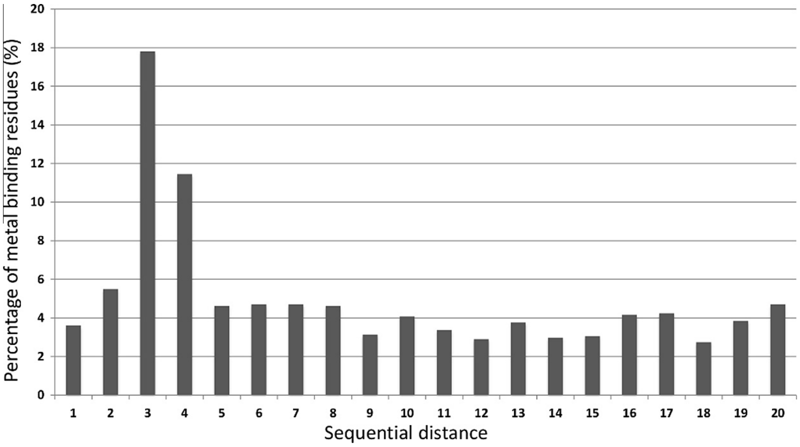

Fig. 1. Frequency of metal binding residues in the vicinity of Cys residues. *X*-axis shows the sequential distance from Cys, while the *Y*-axis is the percentage of metal binding residues at that position. The frequency values are normalized for the first 20 sequential positions, i.e. the sum of all columns adds up to 100.

the centrally located residue in a given test segment was annotated as bound, otherwise "free". When tested, the accuracies of the predictions made for the 19 non-Cys types of residues correlate well with that of performed on Cys, irrespective of the residue type, with values in the range of 67–76% (Supplementary Table 5). This indicates that the analysis of specific sequence environments of residues is rather providing a global information about the whole protein than a specific one of the local segment.

This observation provides an alternative hypothesis, namely that the connection between local sequence composition and Cys oxidation state is indirect: the oxidation state is governed by the subcellular localization which is in turn heavily correlated with amino acid composition. In order to test this alternative hypothesis, we compared the correlation coefficients of the average amino acid compositions of proteins within each cellular localization subclasses. From this analysis, one can see a clear correspondence between the oxidation state/bounding state of cysteines and the cellular localization (Supplementary Table 6). All these observations suggest that taking into account the amino acid compositions alone would be enough to predict the oxidation state of cysteine residues with around 70% accuracy (ignoring membrane proteins since they represent an intermediate type), as the prediction accuracy of the cellular localization of proteins has the same overall accuracy (70%) [33].

We have compared these results with the predictions of SecretomeP2.0 [31]. All proteins were predicted either as extracellular or intracellular and we assigned the oxidation state of cysteins according to the subcellular location of the query protein. The prediction accuracy was 64%, somewhat lower than the one obtained before. This could be acknowledged to a cumulative error partially originating from imperfect predictions of extracellular localization and partially from the imperfect prediction of oxidation state of Cys.

4. Discussion

We developed two disulfide bridge prediction methods in the past [15,17]. Due to the enormous growth of the available training/testing database we presumed that the revision and re-parameterization of the earlier prediction methods would lead to a significantly higher accuracy. However, upon completing the retraining of the method, prediction accuracies remained very similar to the original ones, despite the many fold increase of the input data. We also explored if distinguished positions existed in determining the oxidation state of Cys residues. Comparing the position specific values of the disulfide bonding state potentials only the sequentially 3rd position seemed to be different. However, when we used only this single pair of positions in the prediction, the effect of loss of information due to the limited amount of input surpassed the more specific signal that these positions could provide, and the overall prediction accuracy dropped. This notion seems to be in accordance with the fact that the method indeed predicts protein localization through amino acid composition. The steady rise of accuracies with the widening of the local sequence window used for the prediction shows no position specificity, but the inclusion of more residues leads to a better estimate of the amino acid composition of the whole protein, which enables the effective prediction of localization.

We also explored prediction on randomized sequences removing all local sequence information. Strikingly, the predictive power remained as good as on native sequences. In another test, we applied the prediction method for the other 19 amino acids. Surprisingly, the accuracy of these predictions was comparable to that of Cys despite the fact the different covalent states do not even exist for these residues. Finally, we found strong correlations between the general amino acid composition of proteins, the subcellular localization and the oxidation state of Cys residues. All these results suggest that the prediction methods that try to capture covalent state of Cys residues using their immediate sequence environment are in fact primarily based on the recognition of the subcellular localization, which strongly correlates with residue composition. In that regard the amino acid composition of the Cys flanking 20 residues is a good approximation of the composition of the whole protein. However, the specific local composition of Cys flanking regions can be important in special cases, for example when cysteines take part in metal binding. This notion accounts for the non-random distribution of amino acids in the ±3rd positions.

Due to the high correlation between the redox state of Cys and the cellular localization of proteins (e.g. only 190 of the 1168 extracellular sequences have Cys residues in reduced form), all statistical and machine learning methods essentially recapitulated the subcellular localization of the whole protein instead of the specific redox feature of Cys. In retrospect, the fact that extra- and intracellular localization is strongly connected to residue composition is well known, as a variety of transmembrane prediction methods take advantage of this observation [34]. However, it was not clear that essentially all the discriminatory power of the disulfide bond forming Cys prediction algorithms come from this signal alone and local effects of non-random sequence features do not contribute significantly. This study provides a specific example that the mere accuracy of a given prediction method may not be sufficient to validate an underlying assumption about a biological phenomenon. Of course, our observation must be carefully weighed for each study. There are Cys redox state predictions that do not rely on the assumption that the sequence environment directly encodes a functional signal. For instance, it has been shown that Cys redox state can be predicted with a competitive accuracy using measures of differences in conservation patterns [23]. This relies on the phenomenon that disulfide bonded or functionally important Cys are particularly difficult to replace [10]. Furthermore, the combination of other, especially biochemistry based information, such as mass spectrometry data into the predictions can on one hand further increase prediction accuracy, on the other hand can help to circumvent the issue concerning incorrect biological assumptions.

5. Conclusions

Our results underline the importance of directly challenging a hypothesis from a biological standpoint as we tried in this study. Such precautions in interpreting the high accuracy of a prediction algorithm can help to avoid false biological inference. It is essential to gain a better biological insight into the principles governing the biochemistry of proteins but it should also prove to be very useful to advance the development of more accurate and novel bioinformatics tools.

Acknowledgements

We thank Drs. Zs. Dosztányi, Cs. Magyar and G. E. Tusnády for the critical reading of the manuscript. This work was supported by research grants OTKA NK 100482 and NIH GM096041 and NIH U54GM094662.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fob.2014.03.003.

References

[1] Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. Biochemistry 13 (2), 222–245.

[2] Finkelstein, A.V. and Ptitsyn, O.B. (1971) Statistical analysis of the correlation among amino acid residues in helical, beta-structural and non-regular regions of globular proteins. J. Mol. Biol. 62 (3), 613–624.
[3] McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server. Bioinformatics 16 (4), 404–405.
[4] Rost, B., Yachdav, G. and Liu, J. (2004) The PredictProtein server. Nucl. Acids Res. 32 (suppl 2), W321–W326.
[5] Fiser, A., Dosztanyi, Z. and Simon, I. (1997) The role of long-range interactions in defining the secondary structure of proteins is overestimated. Comput. Appl. Biosci. 13 (3), 297–301.
[6] Gromiha, M.M. and Selvaraj, S. (2004) Inter-residue interactions in protein folding and stability. Prog. Biophys. Mol. Biol. 86 (2), 235–277.
[7] Cox, J.P. et al. (1993) Dissecting the structure of a partially folded protein. Circular dichroism and nuclear magnetic resonance studies of peptides from ubiquitin. J. Mol. Biol. 234 (2), 483–492.
[8] Chuang, C.C. et al. (2003) Relationship between protein structures and disulfide-bonding patterns. Proteins 53 (1), 1–5.
[9] Gupta, A., Van Vlijmen, H.W. and Singh, J. (2004) A classification of disulfide patterns and its relationship to protein structure and function. Protein Sci. 13 (8), 2045–2058.
[10] Rubinstein, R. and Fiser, A. (2008) Predicting disulfide bond connectivity in proteins by correlated mutations analysis. Bioinformatics 24 (4), 498–504.
[11] Singh, R. (2008) A review of algorithmic techniques for disulfide-bond determination. Brief Funct. Genom. Proteom. 7 (2), 157–172.
[12] Czaplewski, C. et al. (2004) Prediction of the structures of proteins with the UNRES force field, including dynamic formation and breaking of disulfide bonds. Protein Eng. Des. Sel. 17 (1), 29–36.
[13] Vonderviszt, F., Matrai, G. and Simon, I. (1986) Characteristic sequential residue environment of amino acids in proteins. Int. J. Pept. Protein Res. 27 (5), 483–492.
[14] Cserzo, M. and Simon, I. (1989) Regularities in the primary structure of proteins. Int. J. Pept. Protein Res. 34 (3), 184–195.
[15] Fiser, A. et al. (1992) Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. FEBS Lett. 302 (2), 117–120.
[16] Muskal, S.M., Holbrook, S.R. and Kim, S.H. (1990) Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng. 3 (8), 667–672.
[17] Fiser, A. and Simon, I. (2000) Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics 16 (3), 251–256.
[18] Xu, H. and Freitas, M.A. (2009) MassMatrix: a database search program for rapid characterization of proteins and peptides from tandem mass spectrometry data. Proteomics 9 (6), 1548–1555.
[19] Murad, W. and Singh, R. (2013) MS2DB+: a software for determination of disulfide bonds using multi-ion analysis. IEEE Trans. Nanobiosci. 12 (2), 69–71.
[20] Schilling, B. et al. (2003) MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. J. Am. Soc. Mass Spectrom. 14 (8), 834–850.
[21] Murad, W., Singh, R. and Yen, T.Y. (2011) An efficient algorithmic approach for mass spectrometry-based disulfide connectivity determination using multi-ion analysis. BMC Bioinformat. 12 (Suppl. 1), S12.
[22] Rose, P.W. et al. (2013) The RCSB Protein Data Bank: new resources for research and education. Nucl. Acids Res. 41 (D1), D475–D482.
[23] Fiser, A. and Simon, I. (2002) Predicting redox state of cysteines in proteins. Methods Enzymol. 353, 10–21.
[24] Huang, Y. and Li, Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20 (1), 21–28.
[25] Tamura, T. and Akutsu, T. (2007) Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. BMC Bioinformat. 8, 466.
[26] Nakashima, H. and Nishikawa, K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J. Mol. Biol. 238 (1), 54–61.
[27] Savojardo, C. et al. (2011) Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization. Bioinformatics 27 (16), 2224–2230.
[28] Szeltner, Z. et al. (2004) Concerted structural changes in the peptidase and the propeller domains of prolyl oligopeptidase are required for substrate binding. J. Mol. Biol. 340 (3), 627–637.
[29] Nardai, G. et al. (2000) Reactive cysteines of the 90-kDa heat shock protein, Hsp90. Arch. Biochem. Biophys. 384 (1), 59–67.
[30] Wu, C.H. et al. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucl. Acids Res. 34 (suppl 1), D187–D191.
[31] Bendtsen, J.D. et al. (2004) Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng. Des. Sel. 17 (4), 349–356.
[32] Imai, K. and Nakai, K. (2010) Prediction of subcellular locations of proteins: where to proceed? Proteomics 10 (22), 3970–3983.
[33] Casadio, R., Martelli, P.L. and Pierleoni, A. (2008) The prediction of protein subcellular localization from sequence. A shortcut to functional genome annotation. Brief Funct. Genom. Proteom. 7 (1), 63–73.
[34] Tusnady, G.E. and Simon, I. (2001) The HMMTOP transmembrane topology prediction server. Bioinformatics 17 (9), 849–850.