Beilstein-Institut

Experimental Standard Conditions of Enzyme Characterization
September 12th – 16th, 2011, Rüdesheim/Rhein, Germany

Metal Binding Sites in Proteins

Vladimir Sobolev[*], Ronen Levy, Mariana Babor and Marvin Edelman[#]

Department of Plant Sciences, Weizmann Institute of Science, Rehovot, Israel

E-Mail: *vladimir.sobolev@weizmann.ac.il and #marvin.edelman@weizmann.ac.il

Received: 22nd February 2012/Published: 15th February 2013

Abstract

Metal ions play a critical role in living systems. About one third of proteins need to bind metal for their stability and/or function. In this review, current sequence based and structure based methods for metal binding site prediction will be presented, with emphasis on the CHED and SeqCHED methods of prediction from apo-protein structures and protein sequences having homologs (even remote) in the structural protein databank (PDB). Metal binding site prediction will be considered as a step in function assignment for new proteins. Finally, a disproportional association of first and second shell metal binding residues in human proteins with disease-related SNPs will be shown.

Introduction

Biological cells must adapt strict regulatory mechanisms in order to maintain metal homeostasis within the cytoplasm [1]. While metal ions can be utilized in various manners in a biological system, the position of a metal ion in space, its variation in time, and the exact chemical partner with which it interacts (often a protein) have been selected by the demands of evolution [2].

Metal ions are required for a great variety of purposes in proteins and are present in more than one third of protein structures investigated [3, 4]. Metals increase the structural stability of the protein in the proper conformation required for biological function. A metal ion can serve as a cross-linking agent, since metal ions usually bind through several interactions with amino acid side chains [2]. In addition, metals can be directly involved in the chemical reactions catalysed by an enzyme. They can serve as redox centres for catalysis (e. g., haem-iron centres) or as electrophilic reactants in catalysis [5]. Metals can help to bring reactive groups into the correct orientation for reaction.

Furthermore, metals can play a regulatory role in proteins. This includes a role in signal transduction, in controlling the architecture of protein complexes, and in redox-active metal sites, where the binding and release of the metal is under redox control [6]. Metals have several valence states, which depending on their ligands can lie close in energy. As a consequence, the metal can be switched from state to state upon binding to a protein, resulting in considerable protein changes [2]. Because of the above, it is important to be able to predict metal binding site based on sequence and/or structural information.

Sequence-based Metal Binding Site Predictions

One of the approaches [7], taken to harvest sequence information, systematically determines all possible metal-binding signatures present in the Protein Data Bank. These signatures, termed MBP (Metal Binding Patterns), include the binding residues and their spacing along the sequence. The method was applied to copper proteins, and a library of metal binding patterns was built. Each MBP is used together with the primary sequence of the corresponding metalloprotein to browse any ensemble of sequences of interest. The level of confidence of this method is variable, ranging between 50% to over 90%, depending on the lengths of the local alignments identified around each binding residue. As this work was applied only to copper, it is not clear to what extent it is applicable to other metals. Moreover, a limitation of this work is that it requires identification of conserved spacing patterns between binding residues and these spacings are not conserved in all cases. Hence, it is not possible to search for a binding residue that is far away in sequence from other binding residues, since the exact spacing can vary greatly among sequences. In another study [8], multiple sequence alignment, entropy (residue conservation) and relative weight of gapless matches were obtained, and the correlation between nearby residues was modelled by support vector machine semi-pattern predictors.

Another algorithm [5] takes subsequences of proteins as input, under the assumption that metal binding residues are influenced by the surrounding environment in nature. The amino acid at the centre of the fragment is the target amino acid, whereas the others are the "neighbours". The fragment sequence is encoded to a feature vector, which contains information on the occurrence probability of the amino acid, the propensities of the secondary structure, and the metal-binding propensity of the amino acid. The feature vector is fed into a neural-network learning machine. The learning machine decides whether the target amino acid binds metal or not. This process is repeated by shifting each time one position along the protein sequence, resulting in a new fragment. With this algorithm, binding residues are identified with higher than 90% sensitivity. However, the limitation of this approach is that it predicts metal binding residues rather than metal binding sites. Therefore, it analyses the probability of each putative binding residue individually, instead of taking into consideration the combined context of all residues belonging to one unified site. In some proteins one protein chain can include more than one binding site (for example, 14% of zinc binding sites fall in this category). Thus, the binding residues of different sites can be erroneously intertwined [6].

A third algorithm [9] scans the sequence around the four main residue types involved in metal binding (Cys, His, Asp, Glu; [10 – 12]) using a window of up to 25 residues, physicochemical features (including conservativity) and correlated mutation analysis derived from multiple sequence alignment.

Structure-based Metal Binding Site Predictions

One of the first algorithms [13] is based on the finding that many metal sites in proteins share a common feature: they are cantered in a shell of hydrophilic ligands, surrounded by a shell of carbon-containing groups. Therefore, it is possible to measure the contrast between groups located at the centre of the sphere (more hydrophilic), and groups located at the outer shell (more hydrophobic) within a radius threshold distance. The contrast function is generally maximal when cantered at or near a metal binding site. However, this algorithm also identified regions of high contrast that were not associated with metal binding, such as charged surface residues and buried, positively-charged residues [14].

A second algorithm [15] is designed specifically for Ca^{2+} binding site prediction, since it is based on the finding that the coordination shell of Ca^{2+} ions in proteins contains almost exclusively oxygen atoms supported by an outer shell of carbon atoms. The bond strength contribution of each ligating oxygen in the inner shell can be evaluated, and the sum of such contributions closely approximates the valence of the bound cation. Assuming local neutralization of charges, the bond strength, or bond order, contributed by each oxygen ligand to the ligated cation is the charge of the cation divided by the number of ligands, or the coordination number. When ligands are asymmetrically disposed around the ligating cation, different bonds are expected to have different strengths. Here, the bond-length correlation to bond order, which is also seen in covalent bonding, can be used to estimate the strength of different bonds in structures. When a protein is embedded in a very fine grid of points and an algorithm is used to calculate the valence of each point (representing a potential binding site), a typical distribution of valence values is obtained. However, only a very small fraction of the points have a significantly large valence value. These points share a tendency to cluster around known Ca^{2+} ions, enabling prediction of such sites.

Sodhi *et al.* [16] calculated the likelihood of a given residue to be a metal ligand by considering multiple sequence alignment of homologous proteins as well as approximate structural information. This method, called MetSite, performed satisfactorily for SCOP database superfamilies [17] where large sets of evolutionary related proteins are available. The algorithm was developed considering 190, 18, 11 and 49 superfamilies for Zn, Fe, Cu and Mn, respectively, while valuation of performance was applied to five, four, one and one cases, respectively. As with Lin *et al.* (2005), MetSite suffers from difficulties to identify the location of a metal binding site by inspecting the distribution of predicted individual residues within the protein structure.

The Fold-X algorithm [18] specializes in predicting the spatial position of a metal in the protein. It uses a library, extracted from the PDB, containing the most common metal spatial positions relative to the corresponding ligating atoms. In the first step, this library is used to search for possible metal positions within the protein structure. Then, an optimization step is performed to find the best position for the predicted metal using the Fold-X force field [19]. The resulting position is used to estimate the energy of binding. At the end, a hydration step to add water ligands is also included. This algorithm is geared to, and performed well in identifying the position of metals in holo forms.

FEATURE, a machine learning method based on a Bayesian classifier was used to identify zinc and calcium binding sites in proteins [20, 21]. This method uses many averaged biochemical and biophysical features in six concentric spherical shells around a suspected site. Shell features include number of atoms, Van der Waals volume, hydrophobicity, solvent accessibility, the presence of different oxygens, nitrogens, carbons and sulphur atoms, amino acid residues, and charges. Similar to Fold-X, FEATURE predicts the position of metal ions within the predicted binding site.

CHED Metal Binding Site Prediction

As mentioned previously, it is well established that four residues: Cys (C), His (H), Glu (E) and Asp (D) (referred to as “CHED” by Babor *et al.* [22]), are the most common amino acids forming soft metal binding sites [10 – 12]. The CHED prediction algorithm [22] is composed of two steps. Step 1 is based on a statistical comparison of holo and apo structure pairs, which showed that at most one ligand side chain reorients upon metal binding [23, 24]. In this step (Fig. 1), the algorithm searches for a 3D constellation of three amino acid residues, whose metal-ligating atoms satisfy distance criteria and where at most one side chain has rotated among the three residues. A binding site is defined as a single triad, or multiple triads that share at least one residue between two or more of them. The second step involves filtration and eliminates false positives. A “mild” filter was created based on the observation that sites composed of a large number of triads tend to be true. Therefore, in cases where a site is found to contain at least five triads, all other putative sites with three or fewer triads are discarded. This filtration deletes about 10% of metal binding sites in apo proteins, yielding a sensitivity (percentage of correctly predicted experimentally known metal binding sites) of 90%. However, among binding sites predicted, 38% proved to be false positives, yielding a selectivity (percentage of correct binding sites among all predicted) of 62%.

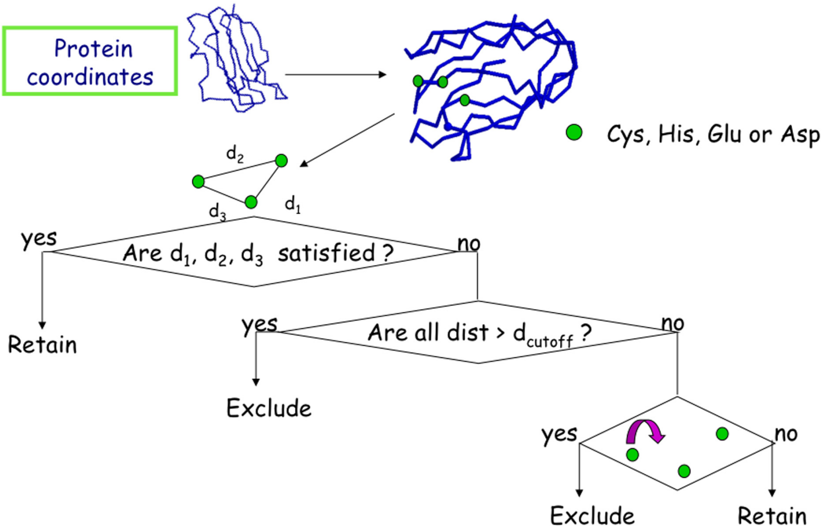

Figure 1. Schematic presentation of step 1 in the CHED algorithm. All possible sets of three amino acid residues (triads) from the four CHED residues, whose collective distances of Cβ atoms are less than 13.0 Å were retrieved. A triad was retained if distances d_1, d_2 and d_3 among ligand atoms from separate CHED residues satisfied individual cutoff criteria. These cutoff values were chosen by analyzing a large (over 1000 sites), redundant set of holo forms and refined using available apo structures. In addition, if one or two out of the three inter-ligand distances were not initially satisfied, alternative side chain conformations of the relevant residues were built, one at the time, using a backbone-independent rotamer library [25]. If no clashes were eventually observed, and d_1, d_2 and d_3 now satisfied the cutoff distances, then the built up triad was retained.

To increase selectivity, a "stringent" filter was created using a decision tree with the following features: number of times a residue of a potential binding site is selected (since a specific residue can belong to more than one initial 'binding site' before joining them together); proportions between C, H, E, D amino acids of a potential binding site; number of sites predicted for the protein; residue sequence entropy; hydrogen bond surface areas between the potential binding residues and any of its neighbouring amino acids. Furthermore, a support vector machine classifier was added, which included the above parameters plus the number of triads per predicted site and relative solvent accessible surface. Triads excluded by both the decision tree and support vector machine classifier were removed. Stringent filtration reduced sensitivity to about 70%. Importantly, it upped selectivity to 90%.

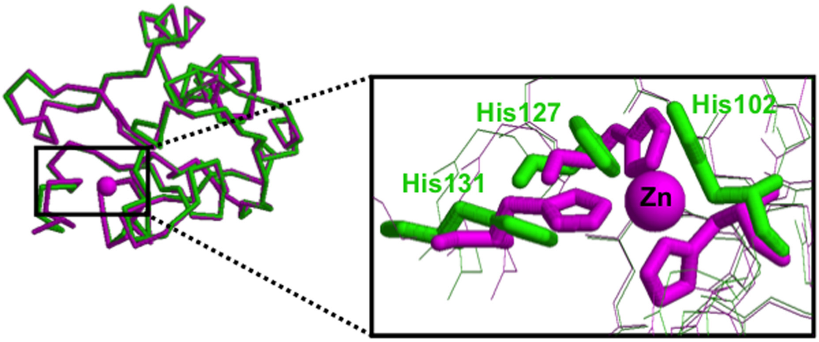

Figure 2. Superimposition of the holo (magenta) and apo (green) colicin E9 DNase domains. The ligand residues for zinc in the apo form were correctly predicted, even though the maximum Cα displacement was equal to 2.45 Å and rearrangement of His102 upon metal ion binding was observed. The coordinates for chains B were taken from PDB entries 2gze (holo form) and 1env (apo form) PDB entries. Binding site residues were found using LPC software [26].

The search procedure for sites has sufficient flexibility built in to often allow for some backbone shifts as well as side chain reorientations. An example is the Colicin E9 DNase domain (Fig. 2). Here the entire binding site was identified successfully in the apo form.

SeqCHED Metal Binding Site Prediction

We developed SeqCHED [27] for prediction of metal binding sites starting with translated DNA sequence data. The method integrates three tools: PsiBlast, SCCOMP and CHED (Fig. 3). PsiBlast [28] identifies statistically significant alignments using a position-specific score matrix that is derived iteratively. The tool was used in a specific manner: target sequences were first subjected to two iterations of PsiBlast against the NCBI non-redundant protein sequence database (NR, http://www.ncbi.nlm.nih.gov/blast/blast_databases.shtml) with a third iteration against PDB to identify structural templates. SCCOMP [29] is a method for side chain modelling. It uses a scoring function including terms for complementarity, excluded volume, internal energy based on roamer probability and solvent accessible surface. The CHED procedure was described above. Table 1 summarizes statistics for SeqCHED prediction. Again, importantly, upon stringent filtration selectivity is higher than 85%.

Table 1. Predictability of transition metal binding sites in modeled structures

PDB template	No. of modeled sites	Sequence identity	Mild filter		Stringent filter	
			% Sensitivity	% Selectivity	% Sensitivity	% Selectivity
Metal containing	223	Target (native)	98	63	93	92
	223	Target (self model)	95	57	84	92
	202	30 – 99%	95	58	84	93
	98	18 – 30%	86	53	85	82
Non-metal containing	143	Target (native)	91	61	76	89
	143	Target (self model)	90	54	67	86
	99	30 – 99%	79	52	49	89
	162	18 – 30%	65	42	33	90

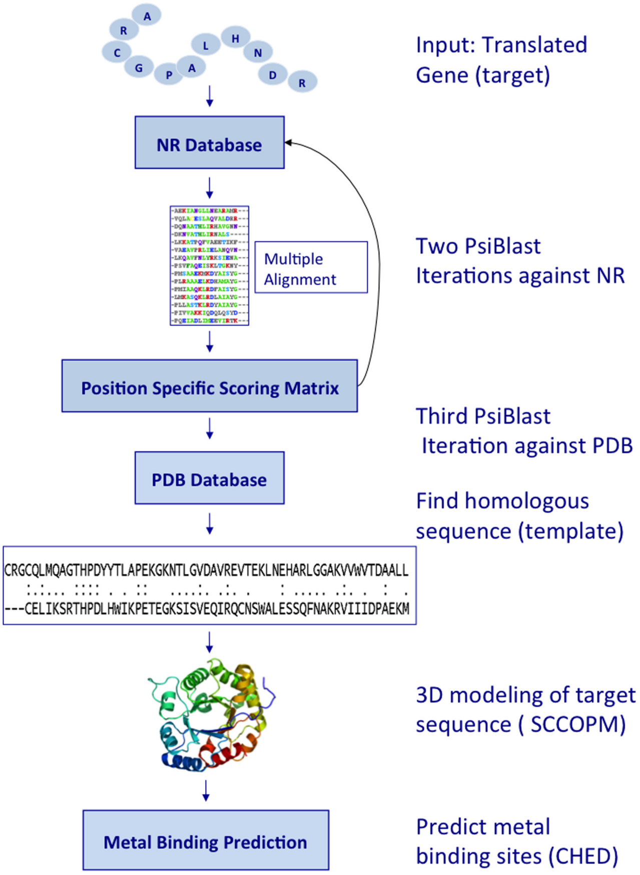
A
R
C
G
P
A
L
H
N
D
R
Input: Translated Gene (target)
NR Database
Multiple Alignment
Two PsiBlast Iterations against NR
Position Specific Scoring Matrix
Third PsiBlast Iteration against PDB
PDB Database
Find homologous sequence (template)
CRGCQLMQAGTHPDYYTLAPEKGKNTLGVDAVREVTEKLNEHARLGGAKVVWVTDAALL
---CELIKSRTHPDLHWIKPETEGKSISVEQIRQCNSWALESSQFNAKRVIIIDPAEKM
3D modeling of target sequence (SCCOPM)
Metal Binding Prediction
Predict metal binding sites (CHED)

Figure 3. Scheme of the SeqCHED procedure. The procedure includes two Psi-Blast iterations against a non-redundant sequence database, a third iteration against the PDB database (to find a structural template), 3D modeling of the target sequence (using SCCOMP for side chain placement), followed by metal binding site prediction using the CHED algorithm.

Linkage between Disease-associated SNPs and Metal Binding Sites

We recently found that mutations associated with diseases (protein variants) are associated with metal binding sites significantly more often than expected [30]. Among the sequences having disease-related single nucleotide polymorphisms (dSNPs), 40% involve mutation of a CHED residue, while for sequences not associated with disease (ndSNPs) the level is 30%. This difference is highly significant and suggests a bias for association of dSNPs with metal binding sites. An analysis of the relation between dSNPs and metal binding sites is presented in Fig. 4. The results demonstrate a clear bias of dSNPs in the immediate vicinity of metal binding sites.

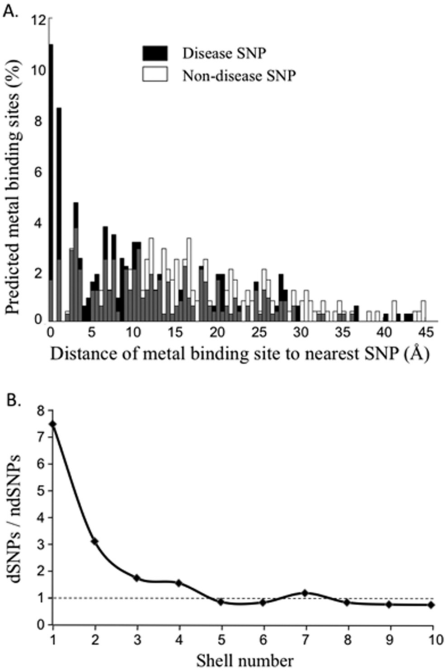

Figure 4. Proximity of dSNPs to metal binding sites (Figure from Levy *et al.* [30]). All proteins containing predicted metal binding sites derived from the Human Polymorphisms and Disease Mutations index were analyzed. The data sets are composed of 237 sequences containing 320 predicted sites with one or more dSNPs, and 184 sequences containing 243 predicted sites with one or more ndSNPs. **A.** The histogram shows the percent distribution of distances between predicted site residues and the nearest dSNP (black bars) or ndSNP (white bars). The overlap between the two distributions is colored gray. The first bar represents the predicted binding site residues (first shell); the bar between 1.0 and 1.5 Å, covalently bound second shell residues; the bars between 2 Å and 4.5 Å, non-covalently bound second shell residues. Bars at greater distances represent residues in successively remote shells. **B.** The normalized ratio between the number of all dSNPs and ndSNPs was obtained for 10 successive shells. A clear differential between the number of dSNPs and ndSNPs can be seen for the first and second shells, and to a lesser extent, for the third and forth shells. The curve reaches a plateau (dashed line) at the fifth shell.

References

[1] Coombs, J.M. and Barkay, T. (2005) New findings on evolution of metal homeostasis genes: Evidence from comparative genome analysis of bacteria and archaea. *Appl. Environ. Microbiol.* **71:**7083 – 7091.
doi: http://dx.doi.org/10.1128/AEM.71.11.7083-7091.2005.

[2] Williams, R.J.P (1985) The symbiosis of metal and protein functions. *Eur. J. Biochem.* **150:**213 – 248.
doi: http://dx.doi.org/10.1111/j.1432-1033.1985.tb09013.x.

[3] Ibers, J.A. and Holm, R.H. (1980) Modeling coordination sites in metallobiomolecules. *Science* **209:**223 – 235.
doi: http://dx.doi.org/10.1126/science.7384796.

[4] Tainer, J.A., Roberts, V.A., and Getzoff, E.D. (1992) Protein metal-binding sites. *Curr. Opin. Biotechnol.* **3:**378 – 387.
doi: http://dx.doi.org/10.1016/0958-1669(92)90166-G.

[5] Lin, C.T., Lin, K.L., Yang, C.H., Chung, I.F., Huang, C.D., Yang, Y.S. (2005) Protein metal binding residue prediction based on neural networks. *Int. J. Neur. Syst.* **15:**71 – 84.
doi: http://dx.doi.org/10.1142/S0129065705000116.

[6] Maret, W. (2005) Fluctuations of cellular, available zinc modulate insulin signaling via inhibition of protein tyrosine phosphatases. *J. Trace Elem. Med. Biol.* **19:**7 – 12.
doi: http://dx.doi.org/10.1016/j.jtemb.2005.02.003.

[7] Andreini, C., Bertini, I., Rosato, A. (2004) A hint to search for metalloproteins in gene banks. *Bioinformatics* **20:**1373 – 1380.
doi: http://dx.doi.org/10.1093/bioinformatics/bth095.

[8] Passerini, A., Andreini, C., Menchetti, S., Rosato, A., Frasconi, P. (2007) Predicting zinc binding at the proteome level. *BMC Bioinf.* **8**, Article Number 39.

[9] Shu, N., Zhou T., Hovmoller S. (2008) Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* **24:**775 – 782.
doi: http://dx.doi.org/10.1093/bioinformatics/btm618.

[10] Alberts, I.L., Nadassy, K., Wodak, S. J. (1998) Analysis of zinc binding sites in protein crystal structures. *Prot. Sci.* **7:**1700 – 1716.
doi: http://dx.doi.org/10.1002/pro.5560070805.

[11] Auld, D. S. (2001) Zinc coordination sphere in biochemical zinc sites. *Biometals* **14:**271 – 313.
doi: http://dx.doi.org/10.1023/A:1012976615056.

[12] Dudev, T., Lin, Y.L., Dudev, M., Lim, C. (2003) First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. *J. Amer. Chem. Soc.* **125:**3168 – 3180.
doi: http://dx.doi.org/10.1021/ja0209722.

[13] Yamashita, M.M., Wesson, L., Eisenman, G., Eisenberg, D. (1990) Where metal-ions bind in proteins. *Proc. Natl. Acad. of Sci. U.S.A.* **87:**5648 – 5652.
doi: http://dx.doi.org/10.1073/pnas.87.15.5648.

[14] Gregory, D.S., Martin, A.C., Cheetham, J.C., Rees, A.R. (1993. The prediction and characterization of metal binding sites in proteins. *Prot. Eng.* **6:**29 – 35.
doi: http://dx.doi.org/10.1093/protein/6.1.29.

[15] Nayal, M. and DiCera, E. (1994) Predicting Ca^{2+}-binding sites in proteins. *Proc. Natl. Acad. of Sci. U.S.A.* **91:**817 – 821.
doi: http://dx.doi.org/10.1073/pnas.91.2.817.

[16] Sodhi, J.S., Bryson, K., McGuffin, L.J., Ward, J.J., Wernisch, L., Jones, D.T. (2004). Predicting metal-binding site residues in low-resolution structural models. *J. Mol. Biol.* **342:**307 – 320.
doi: http://dx.doi.org/10.1016/j.jmb.2004.07.019.

[17] Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:**536 – 540.
doi: http://dx.doi.org/10.1016/S0022-2836(05)80134-2.

[18] Schymkowitz, J.W., Rousseau, F., Martins, I.C., Ferkinghoff-Borg, J., Stricher, F., Serrano, L. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proc. Natl. Acad. Sci. U.S.A.* **102:**10147 – 10152.
doi: http://dx.doi.org/10.1073/pnas.0501980102.

[19] Guerois, R., Nielsen, J.E., Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320:**369 – 387.
doi: http://dx.doi.org/10.1016/S0022-2836(02)00442-4.

[20] Ebert, J.C. and Altman, R.B. (2008) Robust recognition of zinc binding sites in proteins. *Prot. Sci.* **17:**54 – 65.
doi: http://dx.doi.org/10.1110/ps.073138508.

[21] Liu, T. and Altman R.B. (2009) Prediction of calcium-binding sites by combining loop-modeling with machine learning. *BMC Struc. Biol.* **9:**72.
doi: http://dx.doi.org/http://doix.org/10.1186/1472-9-72.

[22] Babor, M., Gerzon, S., Raveh, B., Sobolev, V, Edelman, M. (2008) Prediction of transition metal-binding sites from apo protein structures. *Proteins* **70:**208 – 217. doi: http://dx.doi.org/10.1002/prot.21587.

[23] Babor, M., Greenblatt, H.M., Edelman, M., Sobolev, V. (2005) Flexibility of metal binding sites in proteins on a database scale. *Proteins* **59:**221 – 230. doi: http://dx.doi.org/10.1002/prot.20431.

[24] Edelman, M., Babor, M., Levy, R., Sobolev, V. (2008) Metalloproteins: Structure, conservation and prediction of metal binding sites. In *Structural proteomics and its impact on the life sciences*. eds. Sussman, J.L. and Silman, I., pp. 181 – 205.

[25] Dunbrack, R.L., Jr., and Cohen, F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Prot. Sci.* **6:**1661 – 1681. doi: http://dx.doi.org/10.1002/pro.5560060807.

[26] Sobolev V., Sorokine A., Prilusky J., Abola E.E., Edelman M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15:**327 – 332. doi: http://dx.doi.org/10.1093/bioinformatics/15.4.327.

[27] Levy, R., Edelman, M., Sobolev, V. (2009) Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates. *Proteins* **76:**365 – 374. doi: http://dx.doi.org/10.1002/prot.22352.

[28] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25:**3389 – 3402. doi: http://dx.doi.org/10.1093/nar/25.17.3389.

[29] Eyal, E., Najmanovich, R., Mcconkey, B.J., Edelman, M., Sobolev, V. (2004) Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J. Comput. Chem.* **25:**712 – 724. doi: http://dx.doi.org/10.1002/jcc.10420.

[30] Levy, R., Sobolev, V., Edelman, M. (2011) First- and second-shell metal binding residues in human proteins are disproportionately associated with disease-related SNPs. *Hum. Mutation* **32:**1309 – 1318. doi: http://dx.doi.org/10.1002/humu.21573.