Incorporating post-translational modifications and unnatural amino acids into high-throughput modeling of protein structures

Ken Nagata[1,2], Arlo Randall[1,2] and Pierre Baldi[1,2,*]

[1]Department of Computer Science and [2]Institute for Genomics and Bioinformatics, University of California, Irvine, CA, USA

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Accurately predicting protein side-chain conformations is an important subproblem of the broader protein structure prediction problem. Several methods exist for generating fairly accurate models for moderate-size proteins in seconds or less. However, a major limitation of these methods is their inability to model post-translational modifications (PTMs) and unnatural amino acids. In natural living systems, the chemical groups added following translation are often critical for the function of the protein. In engineered systems, unnatural amino acids are incorporated into proteins to explore structure–function relationships and create novel proteins.

Results: We present a new version of SIDEpro to predict the side chains of proteins containing non-standard amino acids, including 15 of the most frequently observed PTMs in the Protein Data Bank and all types of phosphorylation. SIDEpro uses energy functions that are parameterized by neural networks trained from available data. For PTMs, the χ_1 and χ_{1+2} accuracies are comparable with those obtained for the precursor amino acid, and so are the RMSD values for the atoms shared with the precursor amino acid. In addition, SIDEpro can accommodate any PTM or unnatural amino acid, thus providing a flexible prediction system for high-throughput modeling of proteins beyond the standard amino acids.

Availability and implementation: SIDEpro programs and Web server, rotamer libraries and data are available through the SCRATCH suite of protein structure predictors at http://scratch.proteomics.ics.uci.edu/

Contact: pfbaldi@uci.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Post-translational modifications (PTMs) are critical to the function of many proteins in living systems, and understanding their effects at the molecular level is important for both basic and applied research in biology and medicine. To further this understanding, open databases of curated PTM information have been published. For instance, Phospho.ELM (Dinkel *et al.*, 2010) is a publicly available database dedicated to phosphorylation. The database provides the exact positions of experimentally determined phosphorylation sites as well as information on the specific kinases that produce the modifications. Other databases such as PhosphoSitePlus, HPRD and PHOSIDA (Gnad *et al.*, 2011; Hornbeck *et al.*, 2012; Keshava Prasad *et al.*, 2009) include information on additional types of PTMs (e.g. ubiquitination, acetylation, methylation) but are still dominated by phosphorylation data. An automated curator of information on PTMs (Khoury *et al.*, 2011) found in Swiss-Prot (Bairoch and Apweiler, 2000) provides the following summary statistics: there are a total of 82 505 PTMs determined by experimental methods, with the following types having a frequency >1%: phosphorylation, 70.9%; acetylation, 8.2%; N-linked glycosylation, 6.8%; amidation, 3.5%; hydroxylation, 2.0%; methylation, 1.9%; O-linked glycosylation, 1.4%; ubiquitylation, 1.1% and pyrrolidone carboxylic acid, 1.0%.

In addition to methods for curating and organizing existing PTM data, there are also methods for predicting which sites are modified in sequences with unknown PTM status. These methods typically use supervised machine learning, statistical and motif-based approaches to predict sites of phosphorylation (Blom *et al.*, 1999; Wan *et al.*, 2008), acetylation (Li *et al.*, 2009), glycosylation (Hamby and Hirst, 2008; Julenius *et al.*, 2005; Li *et al.*, 2006), sumoylation (Ren *et al.*, 2009; Xu *et al.*, 2008) and less common types of PTMs as well (Plewczynski *et al.*, 2012). Some of these methods predict both the specific phosphorylated sites and the specific kinases responsible for the phosphorylation (Blom *et al.*, 2004; Kim *et al.*, 2004; Obenauer *et al.*, 2003).

In contrast with these approaches, the fundamental problem of predicting accurate three-dimensional (3D) models of PTMs in proteins has been largely ignored. None of the widely used or recently published side-chain prediction methods that are free for academic research (Hartmann *et al.*, 2007; Krivov *et al.*, 2009; Liang *et al.*, 2011; Lu *et al.*, 2008; Nagata *et al.*, 2012; Zhichao *et al.*, 2011) are capable of incorporating PTMs or unnatural amino acids into their predictions. The widely used template-based modeling software Modeller (Sali and Blundell, 1993) allows for manual creation of custom residues; however, the process for doing so is somewhat cumbersome and not realistic for most Modeller users. One notable exception is the incorporation of non-canonical amino acids into Rosetta (Leaver-Fay *et al.*, 2011) for computational protein–peptide interface design (Renfrew *et al.*, 2012).

The developers of side-chain prediction methods recognize the need for generating accurate models that incorporate PTMs; however, there are a number of practical challenges that have stymied progress in this area: (i) there are far less data in the Protein Data Bank (PDB) (Berman *et al.*, 2002) for PTM

*To whom correspondence should be addressed.

residues than native residues for building rotamer libraries or developing statistical potentials; (ii) although one-character codes (e.g. A for alanine) work well for efficiently defining protein sequences, it is unfeasible to use one-character codes for all possible PTMs (there are >100 PTMs documented in the literature); (iii) some important modifications (e.g. O-linked glycosylation) correspond to broad classes of chemical structures rather than a unique chemical structure, and each of the possible molecules would need to be uniquely identified; and (iv) modified residues are generally larger and contain more rotatable bonds than their natural counterparts.

Beyond the 20 standard amino acids and their PTMs, there are also other natural or synthetic amino acids that can be incorporated into proteins. Two additional natural amino acids, selenocysteine (Sec,U) and pyrrolysine (Pyl,O), are coded in some species by codons that are usually interpreted as stop codons. Pyrrolysine, for instance, is used by some methanogenic archaea in enzymes used to produce methane. In addition, >40 unnatural amino acids have been incorporated into proteins through synthetic biology projects, often by creating a unique codon (recoding) and a corresponding transfer RNA, to explore protein structure and function and create novel proteins (Wang *et al.*, 2009; Xie and Schultz, 2005). A tool for modeling the side chains of these rare natural or unnatural amino acids would also be desirable.

Thus, despite the challenges described above, we have taken on the problem of rapidly generating reasonably accurate 3D side-chain models of proteins that incorporate amino acids beyond the standard 20 amino acids. In the remainder of this article, the term 'non-standard amino acid' (NSA) is used to refer to any amino acid other than the 20 standard amino acids. This includes PTMs, rare natural amino acids and unnatural amino acids.

2 METHODS

2.1 Training and testing datasets

Because we use machine learning methods to predict the side-chain conformations of NSAs, we first describe our curated datasets. We distinguish the 15 more frequent post-translational modifications (FPTMs) from all the other NSAs because there are far more data available for them in the PDB.

2.1.1 NSA dataset The PDB assigns a three-letter identifier to unique chemical structures. The system is used for standard amino acids as well as other chemical structures (e.g. ligands, NSAs) that have coordinates in PDB files. To curate a set of NSAs observed in protein structures, we started from a set of 1449 chemical structures identified as 'non-standard polymeric components' by the PDB. From this starting set, we first removed molecules that were not amino acids, leaving 614 amino acids after this step. Then, we downloaded all of the PDB structure files that contained one or more of these 614 identifiers, yielding 12 294 PDB files. Next, we checked the integrity of the peptide backbone for each potential NSA. If either peptide bond distance was >1.5Å, a feature typically observed with less-constrained amino acids located at the beginning or tail of a protein, the NSA was excluded from the dataset, leaving 603 NSAs after this step. Next, we excluded any NSA that did not have at least one standard amino acid adjacent to it in the peptide chain. After this step, 549 distinct NSAs contained in 12 045 PDB files remained. The reason for this step was to exclude NSAs observed only in short peptides

composed exclusively of NSAs that are never observed integrated into proteins. Then, we excluded NSAs that have no carbon γ or multiple carbon γs because only amino acids with a single χ_1 angle are considered for the prediction stage. After this step, 459 NSAs contained in 11 543 PDB files remained. Next, we excluded proteins with NSAs with high B-factors (>40) because of the uncertainty in the corresponding conformations. Finally, we removed redundancy at the protein sequence level using a sequence similarity threshold of 30% and set aside the data corresponding to the 15 most frequent PTMs (see next section on FPTMs). The final NSA (non-FPTM) dataset consists of 316 unique NSAs contained in 1308 PDBs files. The NSA (non-FTPM) dataset is used exclusively for estimating the generalization accuracy of SIDEpro (see below).

2.1.2 FPTM dataset Our main criterion for categorizing an NSA as a PTM was that a substructure of the NSA must be one of the standard 20 amino acids. We sought to discover the set of PTMs with sufficient instances in the PDB to allow for training and creating rotamer libraries. For this purpose, we set a threshold of at least 50 instances. We sorted the curated NSA dataset by the total number of times the NSA is observed in the PDB. Multiple occurrences in the same PDB file were counted as unique occurrences. After ordering the dataset, we observed that there were 15 NSAs with >50 occurrences, and all of them were PTMs according to our definition. Table 1 shows the chemical structures of the PTMs and their precursor standard amino acids (e.g. tyrosine is the precursor of phosphotyrosine) using the PDB atom naming scheme to label individual atoms.

Selenomethionine (MSE) was associated with a particularly large number of PDB files, and thus we selected 500 of them at random. Finally, for each PTM, the corresponding files were split into five equal folds for cross-validation purposes. The total number of PDB files in the FPTM set is 1168. Supplementary Tables S1 and S2 contain summary information for all the NSAs in the final NSA and FPTM datasets, including PDB three-letter codes, SMILES representations and the corresponding list of PDB file names. Supplementary Table S3 contains the original training set for SIDEpro (Nagata *et al.*, 2012).
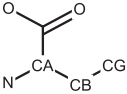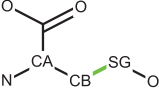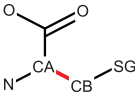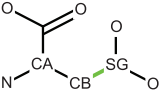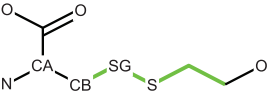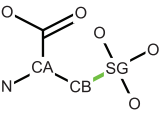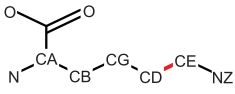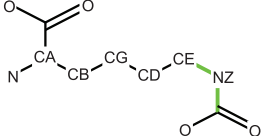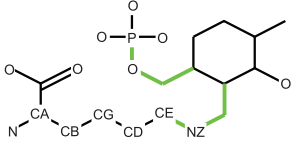
2.2 Building rotamer libraries for NSAs

A *fixed rotamer* is defined by a specific set of χ angles whose values are typically equal to the mean of the values observed in a set of corresponding side-chain conformations that cluster in 3D space. A *flexible rotamer* is defined by both the means and variances of each one of its χ angles. Both types of rotamers are widely used in side-chain conformation prediction, with rigid rotamer libraries (Bhuyan and Gao, 2011; Scouras and Daggett, 2011) generally leading to faster, but slightly less accurate, algorithms than flexible rotamer libraries (Krivov *et al.*, 2009; Lovell *et al.*, 2000; Nagata *et al.*, 2012). Although several rotamer libraries have been published for natural amino acids, only a few exist for NSAs (Gfeller *et al.*, 2012; Renfrew *et al.*, 2012).
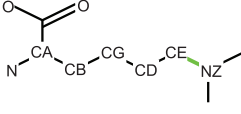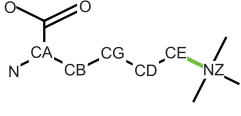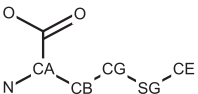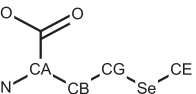
2.2.1 Flexible rotamer library for FPTMs For 14 of the 15 PTMs in this study, the atoms of the precursor amino acid that is being modified are a subset of the atoms in the modified residue. The exception is selenomethionine. Of the 14 FPTMs where the precursor atoms are a subset, 12 introduce new rotatable bonds (i.e. additional χ angles) that must be dealt with. The two FPTMs with proline as the precursor are the exceptions. For instance, serine (SER) has only one χ angle, whereas phosphorylated serine (SEP) has three χ angles.
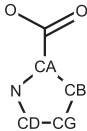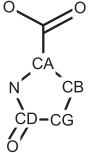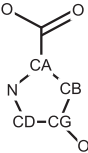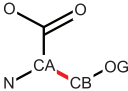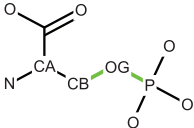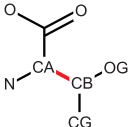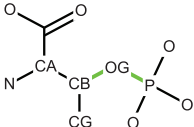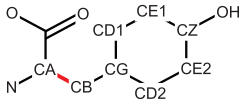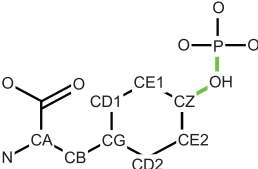
In Table 1, when FPTMs have additional χ angles, the last χ angle of the precursor amino acid is highlighted in red, and the additional χ angles in the FPTM are highlighted in green. For instance, in Table 1 the last χ angle of serine, corresponding to the CA-CB bond, is highlighted in red. The two additional χ angles, corresponding to the CB-OG and OG-P bonds in phosphoserine, are highlighted in green. Note that for phosphotyrosine (PTR), with tyrosine (TYR) as the precursor, the first

Table 1. Frequent post-translational modifications (FPTMs)

Precursor		PTM		
AA	Structure	ID	Name	Structure
ALA		ABA	alpha-Aminobutyric acid	
CYS		CSO	S-Hydroxycysteine	
		CSD	3-Sulfinoalanine	
		CME	S,S-(2-Hydroxyethyl) thiocysteine	
		OCS	Cysteinesulfonic acid	
LYS		KCX	Lysine NZ-carboxylic acid	
		LLP	2-Lysine(3-Hydroxy-2-Methyl-5-Phosphonooxymethyl-pyridin-4-ylmethane)	
		MLY	N-dimethyl-lysine	
		M3L	N-trimethyl-lysine	
MET		MSE	Selenomethionine	

(continued)

Table 1. Continued

Precursor		PTM		
AA	Structure	ID	Name	Structure
PRO		PCA	Pyroglutamic acid	
		HYP	4-Hydroxyproline	
SER		SEP	Phosphoserine	
THR		TPO	Phosphothreonine	
TYR		PTR	O-Phosphotyrosine	

χ angle is treated as the last χ angle because the second (and final) χ angle corresponding to the CB-CG bond is non-rotameric (Shapovalov and Dunbrack, 2011).

The χ angles present in the precursor will be denoted by χ_p, and those that are additional in the modified residue by χ_a. To model the χ angles in FPTM residues that are present in the precursor residues (χ_p), a standard native amino acid rotamer library was used without modification (Shapovalov and Dunbrack, 2011). The additional χ angles in χ_a were handled with a new customized method designed to accommodate cases where only few training instances are available, relative to the case of natural amino acids. For each FPTM type, except LLP and CME, we placed each i-th χ angle (χ_{ai}) in χ_a into one of three angle bins: (0, 120°), (120, 240°) and (240, 360°). We calculated the corresponding means $\mu_{ai}^{r_{ai}}$ and standard deviations $\sigma_{ai}^{r_{ai}}$ where r_{ai} is a rotamer type for χ_{ai}. By assuming that each χ angle is independent, χ_a can be assigned to a maximum of $R_a = 3^{|\chi_a|}$ rotamers (rotamers with zero counts are eliminated).

For symmetric bonds (O-P bonds in LLP, SEP, TPO and PTR; CB-SG bond in OCS; NZ-C bond in KCX; and CE-NZ bond in M3L), because their χ angles are almost constant, we set their mean χ angle to 180° in the rotamer library. The χ angles for PCA are also constant, and thus we set $\chi_1 = 0°$, $\chi_2 = 0°$ and $\chi_3 = 180°$ for PCA. In all these cases, we set the standard deviations to a small default value equal to 10°.

For LLP and CME, because they have many additional χ angles and more possible rotamers, we found that the prediction accuracy is lower comparing with other FPTMs, when using the library defined above. Because of this, we decreased the number of possible rotamers by decreasing the size of the bins. For LLP, the bins are (0, 120°) and (120, 360°) for χ_{a1}; (0, 240°) and (240, 360°) for χ_{a2} and χ_{a4}; (0, 180°) and (180, 360°) for χ_{a3}; and a single bin (0, 360°) for χ_{a5}. For CME, the bins are (0, 180°) and (180, 360°) for χ_{a1} and χ_{a2}; we treated χ_{a3} and χ_{a4} as fixed bonds with values 180 and 300°. These bins were determined from the empirical distribution of χ_a.

We assume that χ_a is dependent on the last χ angle in χ_p, marked in red in Table 1, and referred to as $\chi_{p,last}$. This angle ($\chi_{p,last}$) is associated with one of three bins of equal size 120° as above. For each rotamer of $\chi_{p,last}$, we calculated the rotamer probabilities $p(\mathbf{r_a}|r_{p,last})$, where $\mathbf{r_a}$ is the rotamer types for the additional χ angles, and $r_{p,last}$ is the rotamer type
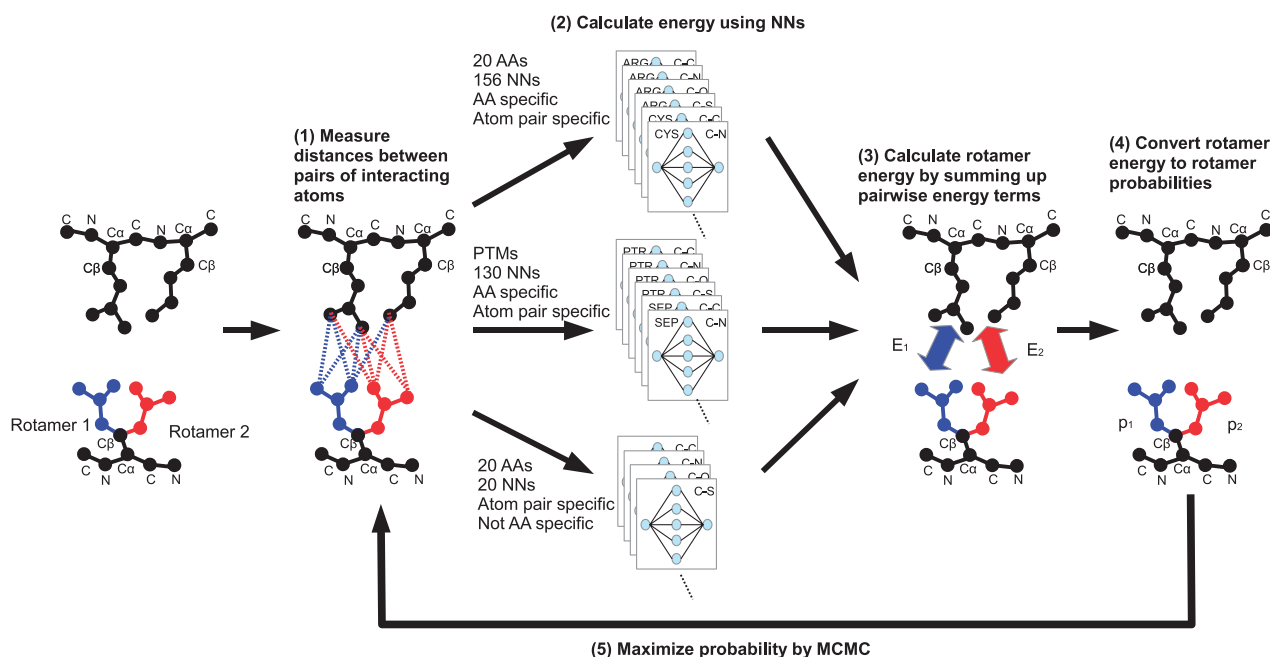
(2) Calculate energy using NNs

(1) Measure distances between pairs of interacting atoms

20 AAs
156 NNs
AA specific
Atom pair specific

PTMs
130 NNs
AA specific
Atom pair specific

20 AAs
20 NNs
Atom pair specific
Not AA specific

Rotamer 1   Rotamer 2

(3) Calculate rotamer energy by summing up pairwise energy terms

E_1 E_2

(4) Convert rotamer energy to rotamer probabilities

p_1 p_2

(5) Maximize probability by MCMC

Fig. 1. Training pipeline

for $\chi_{p,last}$. Because there are R_a rotamers for the additional χ angles in χ_a and R_p rotamers for the precursor residue, the total number of rotamers for a FPTM is $R_a \times R_p$, and the probability of combined rotamer $(\mathbf{r_p}, \mathbf{r_a})$ is $p_p^{\mathbf{r_p}} \times p(\mathbf{r_a}|r_{p,last})$ normalized by the sum of all $R_a \times R_p$ probabilities where $\mathbf{r_p}$ is a rotamer for the precursor residue.

2.2.2 Restricted flexible rotamer library for NSAs (non-FPTM)   Our approach to the generic prediction of NSAs, which do not correspond to FPTMs, treats only the first χ_1 (usually CA-CB) as rotatable and considers the rest of the NSA structure as fixed. We built a general backbone independent flexible rotamer library for the χ_1 angle using the original SIDEpro training dataset (Nagata *et al.*, 2012) (listed in of Supplementary Table S3). First, the χ_1 angles for all natural amino acids (except alanine and glycine, which have no χ_1 angle) in the training set combined (not type specific) were calculated and placed into one of three bins: $(0, 120°)$, $(120, 240°)$ and $(240, 360°)$. The mean and standard deviation of the χ_1 angles for each rotamer bin were calculated. By default, the values of the $\chi_i, i \geq 2$ angles are fixed to those of the original NSA structure. If a user provides multiple structures for a given NSA, SIDEpro automatically builds a uniform rotamer library for $\chi_i, i \geq 2$. For the SIDEpro Web server and downloadable program, we use the COSMOS program (Andronico *et al.*, 2011) for predicting the conformations of small molecules to produce 10 conformations for each NSA that is not an FPTM. The FPTM/NSA rotamer library is given in Supplementary Table S4.

### 2.3   Training energy and prediction

To predict side chains, SIDEpro uses an additive energy function parameterized using a large number of neural networks trained from the data. All the neural networks have similar structure with one input unit corresponding to a distance between a pair of atoms, one hidden layer of hidden units and one linear output unit computing the corresponding energy term.

2.3.1 Neural networks   For natural amino acids, there are 156 neural networks, one per amino acid type and per atom pair type. For instance,

the carbon–carbon neural network for valine computes the 'energy' contribution associated with any pair of carbon atoms, where the first carbon atom is a non-backbone carbon atom in the valine residue under consideration and the second non-valine carbon atom is contained in a spherical neighborhood of 7Å (this carbon atom could be on the backbone or side chain of another residue or in a ligand). These neural networks are part of the original SIDEpro program (Nagata *et al.*, 2012); all the remaining ones are new. For the most FPTMs, there are 130 new neural networks, one per FPTM type and per atom-pair type. Thus, for instance, there is one carbon–carbon neural network for phosphorylated serine. For NSAs (non-FPTMs), we use a more generic approach with 25 neural networks, one per atom pair type. Note that as a slight simplification in all cases, we consider only five atom types (C, H, N, O, S), treating P as if it were C and using H only in the second position of an interaction.

2.3.2 Training   The training pipeline is summarized in Figure 1. For a given protein in the training set with a fixed backbone, we initialize each rotamer to the value closest to the native conformation. Then, we cycle once through each protein in the training set from the C-terminus to the N-terminus. When a given amino acid is being considered, we compute the energy of all its rotamers using the corresponding neural networks. These energies are converted into probabilities and then compared with the native conformation. The mismatch information is used to adjust the weights of the neural networks using Markov chain Monte Carlo methods [see (Nagata *et al.*, 2012) for more details]. For NSAs (non-FPTMs), we use the original SIDEpro training set (Supplementary Materials) of 252 proteins to train generic energy function neural networks using distances between pairs of atoms in all types of natural amino acids. The SIDEpro training set has no redundancy, at the 25% sequence similarity threshold, with the SCWRL4 (Krivov *et al.*, 2009) test set. We experimented with hidden layers of size 5, 10 and 15, noting in general, a degradation of performance with 5, but only small differences between 10 and 15. In all cases, we selected the hidden layer size, which maximizes the cross-validated accuracy (Supplementary Materials). Although we used 5-fold cross-validation to asses the approach, the final production server is trained on the entire data.
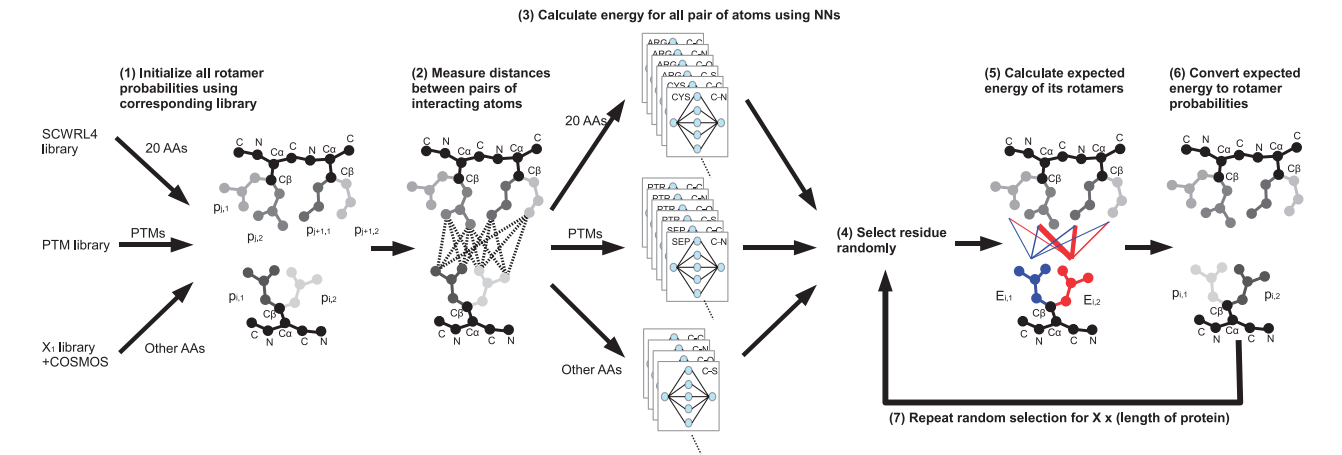
Fig. 2. Prediction pipeline for optimizing all the rotamer probabilities. Once the optimization is completed, final predictions are produced by first selecting the most likely rotamer and then going through a clash reduction algorithm

2.3.3 Prediction The prediction pipeline is summarized in Figure 2. In prediction, we are given a protein with a fixed backbone and possibly, also a set of additional atoms with fixed coordinates, which typically correspond to fixed side chains or atoms in ligand molecules. For the remaining amino acids, we cycle through them in random uniform order without replacement. Each amino acid has its own library of rotamers, and rotamer probabilities as described in Section 2.2. This is true for natural amino acids, FPTMs and other NSAs initialized uniformly over 10 conformations produced by COSMOS. For a given non-fixed amino acid, we compute the expected energy of each one of its rotamers, given all the other fixed atoms, rotamers and rotamer probabilities. These energy values are converted to probabilities, and the corresponding rotamer probability table is updated. The full cycle is repeated six times for each protein.

It is important to note that the neural networks are used only once to compute all the possible energy values because the set of all possible pairwise distances, across all possible rotamer values, does not change during the prediction phase. For the final prediction, we choose the most likely rotamer configuration for each amino acid that is not fixed by the user. Finally, we run the same clash reduction algorithm as in (Nagata *et al.*, 2012).

3 RESULTS

We evaluate the approach using three standard metrics: (i) RMSD for the side chain, which is calculated using the coordinates of the experimental structure, exactly as described in (Krivov *et al.*, 2009); (ii) percentage of side chains where χ_1 is within 40° of the experimental value; and (iii) percentage of side chains where both χ_1 and χ_2 are within 40° of the experimental values.

3.1 Generic energy versus amino acid-specific energy

The generic neural networks and the corresponding energy can first be tested on the 20 natural amino acids and compared with the amino acid-specific neural networks of SIDEpro. Comparison of these two approaches on the SCWRL4 test set, using the SCWRL4 rotamers (Krivov *et al.*, 2009), are reported in Table 2, with a summary for each amino acid of the RMSD, the average χ_1 and average χ_{1+2} and the corresponding *P*-values for a paired *t*-test on the RMSD. For each metric and each

Table 2. AA-specific energy versus generic energy tested on standard amino acids

AA type	AA specific			Generic			*P*-value
	RMSD	χ_1	χ_{1+2}	RMSD	χ_1	χ_{1+2}	
ARG	**2.21**	78.7	64.8	2.23	77.9	64.6	**0.1019**
ASN	**1.04**	85.4	62.5	1.08	83.9	60.9	**0.0002**
ASP	**0.80**	84.9	76.9	0.82	84.3	75.5	**0.0272**
CYS	0.49	90.0	–	**0.47**	90.6	–	0.1755
GLN	1.69	77.3	58.7	1.69	77.4	58.5	0.8869
GLU	**1.46**	**74.1**	**58.0**	1.48	73.7	57.4	**0.0178**
HIS	1.31	88.3	54.8	**1.29**	89.4	53.2	0.3103
ILE	**0.44**	**95.7**	**84.5**	0.46	95.4	82.7	**0.0002**
LEU	**0.53**	**93.9**	**87.6**	0.55	93.4	86.8	**0.0001**
LYS	**1.63**	**79.5**	**66.2**	1.69	77.9	64.9	**<0.0001**
MET	**1.09**	**85.7**	**77.2**	1.11	85.1	75.8	0.1881
PHE	0.67	95.0	92.9	**0.66**	**95.3**	**93.0**	0.1979
PRO	**0.26**	84.7	**81.0**	0.30	**85.3**	79.9	**<0.0001**
SER	**0.75**	73.2	–	0.78	71.5	–	**0.0004**
THR	**0.39**	**90.7**	–	0.40	90.1	–	**0.0239**
TRP	**1.09**	92.7	**84.4**	1.14	92.7	82.7	0.1724
TYR	0.85	94.0	91.4	**0.83**	**94.5**	**91.6**	0.2292
VAL	**0.32**	**93.1**	–	0.34	92.5	–	**0.0057**
ALL	**0.91**	**86.2**	**74.7**	0.93	85.7	73.8	**<0.0001**

amino acid, the best results are shown in bold together with all $P < 0.15$. When all amino acid types are considered as a single large test set, the amino acid-specific neural networks produce slightly more accurate models according to all three metrics with high significance ($P < 0.001$). For 10 amino acid types, the amino acid-specific neural networks perform better than the generic neural networks significantly $P < 0.03$. Note that the generic neural networks produce better results for all three metrics for tyrosine and phenylalanine and for at least one of the three metrics for four other residue types. However, these differences are not statistically significant because there is no amino acid type

Table 3. PTM-specific versus generic energy for frequent PTMs

AA type	Number of pairs	Number of NNs	Number of Pairs/Number of NNs	RMSD Specific	RMSD Generic	*P*-value
ABA	6807	5	1361	**0.99**	1.1433	0.512
CME	4 08 706	15	27 247	**2.83**	2.9194	0.650
CSD	1 08 032	10	10 803	**1.46**	1.598	0.606
CSO	1 73 731	10	17 373	**1.17**	1.18584	0.709
HYP	54 104	10	5410	**1.05**	1.2102	0.411
KCX	13 92 310	15	92 821	**1.72**	1.79402	0.745
LLP	69 02 687	15	4 60 179	**4.08**	5.084	0.002
M3L	3 09 919	10	30 992	**1.71**	1.903	0.195
MLY	16 54 694	10	1 65 469	**2.06**	2.1562	0.128
MSE	81 02 641	10	8 10 264	**1.06**	1.0818	0.105
OCS	45 379	10	4538	0.89	**0.8572**	0.491
PTR	9 26 740	10	92 674	**2.02**	2.2858	0.143
SEP	1 21 541	10	12 154	1.77	**1.7664**	0.993
TPO	10 2497	10	10 250	**1.43**	1.4888	0.255

for which the generic neural networks perform better at a significant level $P < 0.15$. Taken together, these results show overall that (i) as expected, the amino acid-specific neural networks perform better than the generic neural networks on the natural amino acids; (ii) the generic neural networks are not far behind, with RMSDs below 1Å most of the time, and provide reasonable models and a reasonable alternative, with considerably less parameters.

3.2 Prediction of FPTMs

Here we compare the performance of the FPTM-specific neural networks and the generic neural networks for the prediction of FPTMs. One FTPM type, PCA, is excluded from the comparison because it has only one rotamer. We used 5-fold cross-validation on the FPTM datasets. Table 3 shows the average number of atom pairs used for training the FPTM-specific neural networks, the number of neural networks, the ratio of these two numbers and the corresponding cross-validated RMSDs and *P*-values for a paired *t*-test on the RMSDs of each fold. The best RMSD values are in bold together with $P < 0.15$. The number of training atom pairs divided by the number of neural networks provides a rough estimate of the number of examples used for training the neural networks of each FPTM. For four FPTM types (LLP, MLY, PTR and MSE), the FPTM-specific neural networks perform better than the generic neural networks with significance $P < 0.15$. These four types correspond also to the four highest values of the average number of training pairs per neural network, excluding KCX, which has a high *P*-value. For all FPTMs, except OCS and SEP, the specific neural networks perform better, although the difference is small. In the final program, we use the specific energy for all FPTMs.

Table 4 summarizes the cross-validated prediction accuracy results for the FPTMs, grouped according to their precursor amino acid, on the FPTM datasets. Each precursor amino acid is shown in bold together with the corresponding SIDEpro

Table 4. Accuracy for FPTMs and their precursor amino acids

AA type	Count	RMSD(Å) Precursor	All	χ₁ (%)	χ₁₊₂(%)
ALA					
ABA	12	0.99	0.99	61.5	
CYS	**1001**	**0.49**		**90.0**	
CME	16.4	0.86	2.83	75.4	48.0
CSD	16.6	0.56	1.46	91.7	44.0
CSO	44.8	0.64	1.17	88.0	57.1
OCS	17.2	0.60	0.89	86.4	81.6
LYS	**3901**	**1.63**		**79.5**	**66.2**
KCX	14	1.20	1.72	91.3	66.3
LLP	29.4	1.82	4.08	85.9	36.4
M3L	10	1.31	1.71	82.2	70.6
MLY	51.8	1.43	2.06	76.6	64.6
TYR	**2346**	**0.85**		**94.0**	**91.4**
PTR	13.8	1.24	2.02	88.1	79.1
SER	**4107**	**0.75**		**73.2**	
SEP	17.2	0.88	1.77	68.0	39.2
THR	**3790**	**0.39**		**90.7**	
TPO	10.4	0.91	1.43	71.9	66.3
MET	**1410**	**1.09**		**85.7**	**77.2**
MSE	742.6	1.06	1.06	88.0	80.5
PRO	**3233**	**0.26**		**84.7**	**81.0**
HYP	45.8	0.87	1.05	78.7	67.1
PCA	15.2	0.43	0.48	100	100

results. The table shows the average number of instances observed in the test set for each of the 15 FPTMs as well as the cross-validated results for the three accuracy metrics (RMSD, χ₁ and χ₁₊₂). Two average RMSD results are presented using: (i) only the atoms in common with the precursor amino acid; and (ii) all the atoms. The former allows for a direct comparison with the accuracy of SIDEpro on the precursor amino acid.

Considering the RMSD metric, and only atoms shared with the precursor, the accuracy of the FPTMs is comparable with the accuracy of SIDEpro on the precursor amino acids. Four PTMs have lower mean RMSD than their precursor: KCX-lysine, M3L-lysine, MLY-lysine and MSE-methionine. When all the atoms in the PTM amino acid are considered, the average RMSD results are significantly higher, as expected because of the increase in size and number of rotatable bonds of each PTM side chain with respect to its precursor amino acid. Considering the χ₁ metric, six of the FPTMs have higher accuracy values than their precursor amino acid: CSD-cysteine, KCX-lysine, LLP-lysine, M3L-lysine, MSE-methionine and PCA-proline. Six other PTMs have χ₁ accuracy that is within 10% of the corresponding precursor amino acid result. Considering the χ₁₊₂ metric, only the PTMs associated with lysine, tyrosine, methionine and proline can be compared. Of the eight corresponding PTMs, where a direct comparison with the precursor atoms can be made, four have higher accuracy values than their precursor: KCX-lysine, M3L-lysine, MSE-methionine and PCA-proline. In short, by multiple metrics, the

Table 5. Accuracy of NSA method on FPTM set

AA type	Count	True structure				COSMOS			
		RMSD(Å)		$\chi_1(\%)$	$\chi_{1+2(\%)}$	RMSD(Å)		$\chi_1(\%)$	$\chi_{1+2(\%)}$
		Precursor	All			Precursor	All		
ABA	60	**1.054**	**1.054**	**60**		1.162	1.162	55	
CME	82	**0.6276**	**0.3462**	86.59	**86.59**	0.6648	4.783	**89.02**	3.659
CSD	83	**0.6829**	**0.782**	**86.75**	**86.75**	1.17	2.387	71.08	12.05
CSO	224	**0.6782**	**0.6404**	85.2	84.75	0.8186	1.892	81.25	8.482
HYP	229	**0.1067**	**0.1331**	**100**	**100**	0.5366	0.5366	95.2	29.26
KCX	70	**0.3971**	**0.5761**	**98.57**	**98.57**	2.151	3.377	74.29	47.14
LLP	147	**1.056**	2.029	**89.12**	**89.12**	1.919	6.709	87.76	42.86
M3L	50	**1.125**	**1.407**	**80**	**80**	2.059	2.535	64	56
MLY	259	**1.044**	**1.259**	**83.01**	**83.01**	2.098	2.777	64.48	52.12
MSE	3713	**0.367**	**0.367**	**85.38**	**85.29**	2.023	2.023	82.01	50.9
OCS	86	**0.5527**	**0.7119**	**91.86**	**91.86**	0.785	1.096	86.05	74.42
PCA	76	**0.08948**	**0.08601**	**100**	**100**	0.7811	0.8356	56.58	55.26
PTR	69	**1.43**	2.019	**82.61**	**82.61**	2.215	3.553	69.57	60.87
SEP	86	**0.8974**	**1.411**	65.12	**65.12**	0.9638	2.054	**66.28**	25.58
TPO	52	0.9129	**1.397**	70.59	**70.59**	**0.9065**	1.793	**75**	63.46

prediction accuracy of SIDepro on the 15 FPTMs is roughly comparable with its accuracy on the natural amino acids.

3.3 Prediction of NSAs

The generic NSA prediction method requires a 3D structure model of the NSA be provided as input, and to test the NSA method with more data, we tested it on both the FPTM and the NSA (non-FPTM) test sets. Structure models are derived from two sources: (i) true structures from the PDB; and (ii) conformations generated by COSMOS (Andronico *et al.*, 2011). Results obtained using true structures do not reflect what can be expected from prediction in a realistic setting, but rather provide a sense of the limits of the methods. In true prediction mode, the structure of the NSAs must be generated by a small molecule structure predictor.

Table 5 reports the results of the generic NSA prediction method on the FPTM set, when the FPTM amino acids are treated as non-standard. The best results for each metric and each FPTM are in bold. In this experiment, for each modified amino acid, we use a single predicted structure obtained with COSMOS. As shown below, further improvements can be obtained by using multiple predicted structures. As expected, with a few exceptions, when the true structures are used as input the resulting models are more accurate than when predicted structures are used as input. Overall, the predicted structures lead to reasonable performance, given the complexity of the problem and the high-throughput nature of the approach. In all cases using predicted structure leads to RMSD values that are always < 2.5Å on the shared atoms.

Finally, Table 6 summarizes the results obtained on the NSA (non-FPTM) test set. For this experiment, we compare the results obtained using the true structure from the PDB, a single predicted structure and multiple (10) predicted structures as

Table 6. Accuracy of NSA method on NSA (non-FPTM) test set

Default structures	RMSD(Å)	$\chi_1(\%)$
Single predicted structure	3.54	56.39
10 predicted structures	3.08	65.30
True structure	**1.75**	**66.63**

structural models for the NSAs. The best results for each metric are shown in bold. As expected, using the true structure provides the most accurate results, with an average RMSD of 1.75Å and a χ_1 of 66.63%. Using multiple predicted structures helps improve the performance. For instance, the average RMSD improves from 3.54 to 3.08 A, a value that is reasonable, given the high-throughput nature of the approach and the complexity and variability of NSAs, but requiring further refinements for high-precision tasks. In terms of the χ_1 metric, using 10 structures improves the performance from 56.39 to 65.30%, a value close to the performance obtained using the PDB structures.

4 CONCLUSION

The strength of SIDEpro is that it uses the wealth of data in the PDB to learn energy functions, parameterized by neural networks, to model and predict protein side-chain conformations. In this study, we have extended the capabilities of SIDEpro to PTMs and NSAs.

For natural amino acids and FPTMs, when sufficient training examples are available, SIDEpro uses amino acid-specific energy functions. For all other PTMs and NSAs, SIDEpro uses a generic energy function. To flexibly accommodate for any NSA,

SIDEpro allow users to provide 3D structures of NSAs to be incorporated into SIDEpro models. Alternatively, the COSMOS (Andronico *et al.*, 2011) program is used to predict these structures, and any other similar program [e.g. OpenBabel (O'Boyle *et al.*, 2011)] can be used for the same purposes. The generic neural networks, trained on all possible pairs of atom types agnostic of residue type, are used to score the atom–atom interactions for these NSAs. Naturally, as more data on NSAs become available in the PDB, it will be possible to further expand the set of specific energy functions, thereby increasing the accuracy of the program over time. As demonstrated here for some of the NSAs, accuracy can also be improved by increasing the number of 3D samples produced by COSMOS, at the expense of time.

Finally, SIDEpro is to be used in protein structure prediction and engineering projects for the rapid prediction of side-chains conformations in high-throughput mode, or to provide good starting points for molecular or quantum mechanics simulations of side-chain atoms, for both standard and non-standard amino acids.

ACKNOWLEDGMENTS

REFERENCES

Andronico,A. *et al.* (2011) Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. *J. Chem. Inf. Model.*, **51**, 760–766.

Bairoch,A. and Apweiler,R. (2000) The swiss-prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Berman,H. *et al.* (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Bhuyan,M.S. and Gao,X. (2011) A protein-dependent side-chain rotamer library. *BMC Bioinformatics*, **12** (Suppl. **14**), S10.

Blom,N. *et al.* (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

Blom,N. *et al.* (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Dinkel,H. *et al.* (2010) Phospho.ELM: a database of phosphorylation sitesupdate 2011. *Nucleic Acids Res.*, **39**, D261–D267.

Gfeller,D. *et al.* (2012) Expanding molecular modeling and design tools to non-natural sidechains. *J. Comput. Chem.*, **33**, 1525–1535.

Gnad,F. *et al.* (2011) Phosida 2011: the posttranslational modification database. *Nucleic Acids Res.*, **39** (Suppl. **1**), D253–D260.

Hamby,S. and Hirst,J. (2008) Prediction of glycosylation sites using random forests. *BMC Bioinformatics*, **9**, 500.

Hartmann,C. *et al.* (2007) Irecs: a new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci.*, **16**, 1294–1307.

Hornbeck,P.V. *et al.* (2012) Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.

Julenius,K. *et al.* (2005) Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology*, **15**, 153–164.

Keshava Prasad,T.S. *et al.* (2009) Human protein reference database2009 update. *Nucleic Acids Res.*, **37** (Suppl. **1**), D767–D772.

Khoury,G.A. *et al.* (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.*, **1**, 90.

Kim,J.H. *et al.* (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.

Krivov,G.G. *et al.* (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, **77**, 778–795.

Leaver-Fay,A. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, **487**, 545–574.

Li,S. *et al.* (2006) Predicting O-glycosylation sites in mammalian proteins by using SVMs. *Comput. Biol. Chem.*, **30**, 203–208.

Li,S. *et al.* (2009) Improved prediction of lysine acetylation by support vector machines. *Protein Pept. Lett.*, **16**, 977–983.

Liang,S. *et al.* (2011) Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*, **27**, 2913–2914.

Lovell,S.C. *et al.* (2000) The penultimate rotamer library. *Proteins*, **40**, 389–408.

Lu,M. *et al.* (2008) OPUS-Rota: a fast and accurate method for side-chain modeling. *Protein Sci.*, **17**, 1576–1585.

Nagata,K. *et al.* (2012) Sidepro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. *Proteins*, **80**, 142–153.

Obenauer,J.C. *et al.* (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.

O'Boyle,N. *et al.* (2011) Open babel: an open chemical toolbox. *J. Chemoinform.*, **3**, 33.

Plewczynski,D. *et al.* (2012) AMS 4.0: consensus prediction of post-translational modifications in protein sequences. *Amino Acids*, **43**, 573–582.

Ren,J. *et al.* (2009) Systematic study of protein sumoylation: development of a site-specific predictor of SUMOsp 2.0. *Proteomics*, **9**, 3409–3412.

Renfrew,P.D. *et al.* (2012) Incorporation of noncanonical amino acids into Rosetta and use in computational protein-peptide interface design. *PLoS One*, **7**, e32637.

Sali,A. and Blundell,T. (1993) Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Scouras,A.D. and Daggett,V. (2011) The Dynameomics rotamer library: amino acid side chain conformations and dynamics from comprehensive molecular dynamics simulations in water. *Protein Sci.*, **20**, 341–352.

Shapovalov,M.V. and Dunbrack,R.L. Jr. (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, **19**, 844–858.

Wan,J. *et al.* (2008) Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. *Nucleic Acids Res.*, **36**, e22.

Wang,Q. *et al.* (2009) Expanding the genetic code for biological studies. *Chem. Biol.*, **16**, 323.

Xie,J. and Schultz,P.G. (2005) Adding amino acids to the genetic repertoire. *Curr. Opin. Chem. Biol.*, **9**, 548–554.

Xu,J. *et al.* (2008) A novel method for high accuracy sumoylation from protein sequences. *BMC Bioinformatics*, **9**, 8.

Zhichao,M. *et al.* (2011) Rasp: rapid modeling of protein side-chain conformations. *Bioinformatics*, **27**, 3117–3122.