Critical assessment of side chain conformation prediction in modelling of single point amino acid mutation

Anna Marabotti[§] and Angelo Facchiano

Laboratory of Bioinformatics and Computational Biology, Institute of Food Science, National Research Council, Via Roma 64 – 83100 Avellino (Italy)

[§]Corresponding Author. E-mail: amarabotti@isa.cnr.it

Full description of Methods

Benchmark for the evaluation

The structures of the wild type phage T4 lysozyme (Matsumura et al. 1989) and of the "pseudo-wt" lysozyme carrying the mutations C54T and C97A (Nicholson et al. 1991), obtained by X-ray methods and deposited in the PDB database (Berman et al. 2007) (PDB codes: 3LZM and 1L63, respectively) were chosen as starting points to model the mutants included in the benchmark. The choice of these two reference structures among those available was made on the basis of crystallization conditions and absence of phenomena that could alter side chain conformation in an unpredictable way (for example, high temperature, different solvents, unnatural pH conditions, and so on).

Among the lysozyme mutants whose structure was obtained by X-ray methods, we included in our benchmark those carrying only a single mutation (for "pseudo-wt", in addition to the two cited mutations). Their resolutions range between 1.65 and 2.60 Å. All available mutations were taken into account, with the exception of those involving proline, since this residue strongly affects backbone conformation, making the modification of this residue more difficult to manage computationally with the selected programs. Structures with ligands were not included in the benchmark.

The PDB codes of structures of the mutants obtained from wild type lysozyme included in the benchmark are: 149L; 150L; 1D9W; 1DYC; 1DYE; 1L00; 1L02; 1L06; 1L07; 1L08; 1L09; 1L10; 1L12; 1L13; 1L14; 1L15; 1L16; 1L17; 1L18; 1L19; 1L20; 1L21; 1L22; 1L23; 1L33; 1L34; 1L37; 1L38; 1L44; 1L45; 1L46; 1L47; 1L48; 1L52; 1L53; 1L60; 1L98; 1L99.

The PDB codes of structures of the mutants obtained from pseudo-wild type lysozyme included in the benchmark are: 107L; 108L; 109L; 110L; 111L; 113L; 114L; 115L; 118L; 119L; 120L; 122L; 123L; 125L; 126L; 127L; 128L; 129L; 130L; 131L; 137L; 138L; 160L; 161L; 162L; 163L; 164L; 165L; 166L; 171L; 175L; 195L; 196L; 1CTW; 1CU0; 1CU2; 1CU3; 1CU5; 1CU6; 1CUP; 1CUQ; 1CV0; 1CV1; 1CV3; 1CV4; 1CV5; 1CV6; 1CVK; 1G06; 1G07; 1G0G; 1G0J; 1G0K; 1G0L; 1G0M; 1G0P; 1G0Q; 1G1V; 1G1W; 1I6S; 1JQU; 1L54; 1L55; 1L59; 1L61; 1L62; 1L65; 1L66; 1L67; 1L68; 1L77; 1L85; 1L86; 1L87; 1L88; 1L90; 1L91; 1L92; 1L93; 1L94; 1L95; 1LYE; 1LYF; 1LYG; 1LYH; 1LYI; 1LYJ; 1QS5; 1QS9; 1QSB; 1QSQ; 1QT3; 1QT4; 1QT5; 1QT6; 1QT7; 1QT8; 1QTB; 1QTC; 1QTD; 1QTH; 1QTV; 1QTZ; 1QUD; 1QUG; 1TLA; 200L; 216L; 217L; 221L; 224L; 255L.

Side chain modelling

SCWRL3.0 (Canutescu et al. 2003) and SCAP, included in the JACKAL package (Xiang and Honig 2001) were freely downloaded from the related Web servers, whereas NCN (Peterson et al. 2004) was obtained by direct request to the authors. The average computational cost for each mutation was 2" for SCWRL3.0, 3' for SCAP and 27' for NCN on a COMPAQ/HP server ProLiant ML530. The results of a mutation are fully reproducible for all the programs.

Single-point mutants were modelled on the crystallographic structure of wild type or "pseudo wt" lysozymes by using facilities of the repacking program to insert point mutations (for SCWRL3.0 and SCAP), or by editing the PDB file (for NCN), as indicated in the manual provided by the authors. Then, we allowed the programs to rebuild all the side chains of each mutant lysozyme. For SCWRL3.0, no additional settings for the repacking algorithm are present, apart from the choice of side chains to be repacked. For NCN, default options of the program were applied (Peterson et al. 2004). For SCAP, we chose to apply the default options for the force field used (CHARMM), and the choice of the large side chain rotamer library. We applied the option to retain the original side chain conformation as the first initial conformation, treated as a regular rotamer and included in the rotamer library. For the minimization step, we chose to iteratively sample all side chain rotamers until convergence, then to minimize the final lowest-energy conformation by refining the side chain conformation with 2° rotation on each bond to search for lower energy conformations around the rotamer. These options were chosen according to some examples included in the tutorial provided by the authors.

Evaluation of side chain geometry

The structure of each mutant created as described above was compared to that of the corresponding mutant included in the benchmark and obtained by X-ray crystallography. At present, no standardized protocols to assess the performances of these software packages are available, and, although several parameters are of general use, their combination and thresholds change from one paper to another (Marabotti, 2008). In our work, the accuracy of the side chain conformer predictions was assessed in terms of dihedral angle deviation and RMSD value from the crystallographic conformation.

The program "CHI", implemented in the JACKAL package developed by Honig's group and downloadable from the site:
http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Jackal
was used to calculate dihedral angles, used to evaluate χ_1 and χ_{1+2} correctness. A dihedral angle was considered to be correctly predicted when its value was within $\pm$ 20° of that of the native structure. This threshold is the same adopted by SCAP developers to evaluate the performances of their software (Xiang et al. 2001), but is more restrictive than the one adopted to evaluate other program's performances (Canutescu et al. 2003; Peterson et al. 2004). We decided to adopt the most stringent criterion since it is well known that even very small differences could affect protein behavior. We calculated the percentage of correct χ_{1+2} angles only when χ_1 angles were correctly predicted.

RMSD calculation was performed using the McLachlan algorithm (McLachlan 1982) as implemented in the program ProFit v. 2.3.5.1, developed by Dr. A. C. R. Martin and available at the Web site: http://www.bioinf.org.uk/software/profit. We considered only side chain heavy atoms, either excluding or including Cβ atoms, which lower the RMSD value because typically the deviation for this atom is near zero (Peterson et al. 2004). The residue RMSD is calculated as a by-residue RMSD.

To identify the "neighbours" of a mutations, we selected the residues with at least one atom included in a distance of 5 Å from the mutant side chain (this selection was made using the command "Zone" in the module "Subset" of the program InsightII (Version 2000.1, Accelrys, Inc.)).

The solvent exposure of the residues, measured in terms of percentage of solvent accessible surface area (SASA) of the side chain, was calculated with the program NACCESS (Hubbard et al. 1991; Lee and Richards 1971).

Information on B-factors was extracted from the PDB files of the reference structures of the phage T4 lysozyme included in the benchmark. We calculated a mean B-factor value for each side chain, including Cβ atoms.

To evaluate the conservation of size and polarity of the mutants with respect to the wild type structure, we defined as “conserved in size” all the residues with no more than 10% of difference in their volume, taking as reference values those conserved in the Amino Acid Repository of Jena Library of Biological Macromolecules at the Web address: http://www.fli-leibniz.de/IMAGE_AA.html and calculated according to Zamyatin (Zamyatin 1972). We classified Gly, Ala, Val, Leu, Ile, Met, Phe, Trp as “non polar amino acids”, Asn, Cys, Gln, Ser, Thr, Tyr as “polar amino acids”, and Arg, Lys, Glu, Asp, His as “charged amino acids”. All mutations involving amino acids included in the same group were classified as “unchanged polarity”, with the exception of charged amino acids, in which the replacement of Asp/Glu with Arg/Lys/His and *vice versa* was classified as “charge inversion”.