Supporting information for: Equilibrium simulations of proteins using molecular fragment replacement and NMR chemical shifts

Wouter Boomsma [*†‡], **Pengfei Tian** [§‡], **Jes Frellsen** [¶], **Jesper Ferkinghoff-Borg** [‖], **Thomas Hamelryck** [*], **Kresten Lindorff-Larsen** [*] **and Michele Vendruscolo** [**†]

[*]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark, [§]Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen, Denmark, [¶]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, United Kingdom, [‖]Department of Systems Biology, DTU, Søltofts Plads, 221 DK-2800, Kgs. Lyngby, Denmark, and [**]Department of Chemistry, Lensfield Road, CB2 1EW, Cambridge, United Kingdom

Model estimation and analysis

Training set selection RefDB [1] is a database of X-ray and NMR structures with associated chemical shifts that have been re-referenced to be internally consistent, and thereby constitutes an ideal training set for the model in this paper. The version used in this paper was downloaded in April 2011. This dataset was filtered to exclude any proteins which were in the same SCOP superfamily [2] or had the same CATH architecture [3] as proteins used for testing in this paper (see Fig. S4 for the complete list). For each of the remaining set of 1349 structures, sequences of amino acid labels, dihedral angle pairs and chemical shift values were extracted, and the DSSP program [4] was used to assign the secondary structure.

The Ubiquitin simulations were added later, and consequently, Ubiquitin had not been excluded from the original training set. A separate model was therefore prepared for this purpose, using the same criteria as above to exclude any Ubiquitin-related proteins from the dataset (See Fig. S4).

The effect of redundancy in the dataset was probed by experimenting with weighting the individual sequences based on the size of the corresponding protein family, but no apparent effect on model quality was observed, and we therefore used equal weights for all RefDB entries for the final model presented in this paper.

Training procedure The number of parameters of the model, n_p, is a sum of the contributions from the transition matrix $(60-1)$, the torus node (5: κ_1, κ_2, κ_3, μ_1, μ_2), the cis/trans node (1), the secondary structure node (2), the amino acid node (19), and 6 chemical shift nodes (2: μ, σ), multiplied by the number of hidden node states:

$$n_p = 60(59+5+1+2+19+12) = 5880 \qquad [1]$$

The 1349 proteins give rise to 138283 observations of each of the emission nodes, and there is thus more than enough data to reliably estimate the parameters in the model, which is evident from the similar likelihood scores obtained in repeated estimations of the same model (Fig. S1). We used the stochastic EM (SEM) algorithm [5] to estimate the parameters. In each iteration, the procedure consisted of two steps: 1) for each protein in the training set, all hidden nodes were resampled using the forward-backtrack algorithm [6, 7], which assigned the input data for each residue in the training set to a specific hidden node component; 2) the parameters were updated using maximum likelihood as if the model was fully observed. The SEM algorithm has been shown previously to work well for estimating models of this type [6, 5, 8].

The number of hidden node states was determined by training models of different size, and using the Bayesian Information Criterion [9] to select the appropriate model. Since the training process is stochastic, each model was trained five times, and the highest scoring model (with 60 states in our case) was selected for use in this paper (Fig. S1).

Information encoded in the hidden nodes The hidden node states do not have a direct physical interpretation. They are merely a convenient mechanism for encoding the sequential dependencies along the protein chain, using discrete states rather than for instance the continuous (ϕ, ψ) angular values. The hidden node states can be understood as a classification of local structure similar to the classic secondary structure classification, but using 60 states rather than the usual three. Each state corresponds to a particular distribution in (ϕ, ψ) angular space, and to a distribution of amino acids that describes the preferences of particular amino acids to adopt this state. Likewise, they correspond to a particular chemical shift signal that correlates with the associated (ϕ, ψ) distribution. The transition matrix encodes the probability of moving from one state to the other, and is thus similar in spirit to a Zimm-Bragg helix coil transition model [10], but generalized to a higher number of structural states. As an example, we highlight three states in Fig. S3, representing three different secondary structure preferences, and corresponding differences in amino acid and chemical shift distributions.

Analysis of correlation length in the model An eigen-analysis of the hidden node transition matrix can provide an estimate of the correlation length in the model, measured in terms of residues along the chain. The highest eigenvalue will be unity, corresponding to the stationary state of the model, while lower values indicate the slowest decaying states in the model [11]. For the model in this paper, the next-to highest eigenvalue is 0.84, which implies a correlation length of $\tau = 1/(1-0.84) = 6.25$. Fragment libraries will typically contain fragments that are longer than this, which implies that certain longer range local signals are not captured in the current model. A natural topic for future research would be to extend our model to capture these effects, for instance by employing higher order Markov models or multiple layers of hidden nodes.

Simulation setup

Choice of moves We chose Monte Carlo moves similar to a set that has been used successfully in the past to fold peptides and small proteins [12]. The set consists of a single side chain move (uniform proposals of side chain χ angle updates), and two backbone moves: a pivot-like move which alters a single backbone dihedral pair, and a semi-local move which alters a stretch of dihedral angles, but restrains the movement of the endpoint of the stretch [13]. The pivot-move used either unbiased proposals for the dihedral changes (unbiased simulations), or dihedral angles sampled from the CS-TORUS

[†]To whom correspondence should be addressed. Email: wb@bio.ku.dk, mv245@cam.ac.uk
[‡]W.B and P.T contributed equally to this work.

model. As described in the main text, the CS-TORUS supports efficient resampling of entire stretches of dihedral angles using the forward-backtrack algorithm [6, 7]. The choice of altering only a single dihedral angle pair at a time was made to ensure a fair comparison between unbiased and biased simulations. The acceptance rate of standard pivot moves drops when altering many dihedral angle pairs at once, which would penalize the unbiased simulations excessively.

Generalized Ensembles Simulations were conducted using generalized ensembles in the MUNINN software library [14, 15]. Rather than sampling from the Boltzmann (canonical) distribution, generalized ensembles replace the $\exp(-\beta E)$ term with a weight function $w(E)$: $P_C(x) = Z_C^{-1} w(E(x))$. In the multicanonical ensemble, the goal is a uniform distribution over energies, which is obtained by setting the weight function to the inverse of the density of states, $g(E)$. After conducting a simulation in this ensemble, it is possible to reconstruct average properties according to the Boltzmann distribution at a given temperature using a reweighting technique. The multicanonical method can thus be viewed as a method which simultaneously collects statistics at different Boltzmann distributions corresponding to a range of temperatures. For our purpose, it is convenient to rewrite the Boltzmann factor in units of $1/kT_0$: $P_{MC}(x) \propto \exp(-\gamma \frac{1}{kT_0} E(x))$, where T_0 is the temperature at which we wish to extract statistics, and we specify our temperature range in terms of the scaling factor γ. In a biased simulation, we have a factor P_{DBN} arising from the proposal distribution, $P_{MC}(x) \propto \exp(-\gamma \frac{1}{kT_0} E(x)) P_{DBN}(x)$. This factor can be viewed as an implicit energy, but it should be noted that it does not scale with temperature. However, we can set up the simulation such that the bias will cancel out at T_0 by constructing the modified energy $\tilde{E}(x) = E(x) - \ln(P_{DBN}(x))$. At $\gamma = 0$, only the proposal distribution is active, corresponding to a simulation where the only energy is the chemical shift signal. At $\gamma = 1$, the proposal bias fully cancels the corresponding negative bias term in the explicit energy function, corresponding to an ensemble where only the original force field is used. This procedure corresponds to variant five in Table S2, which gives an overview of different simulation strategies that are available when using a probabilistic proposal distribution. For further details, we refer to ref. [16] and [17].

Trajectory analysis The round-trip time is measured based on the Q-factor reaction coordinate, which is related to the fraction of native contacts formed, and has previously been reported as a good reaction coordinate for protein dynamics studies [18, 19, 20]. We use a similar definition of Q-factor as was recently used for folding time calculations in molecular dynamics simulations [21]. In order to find the 'native contacts', we first divide all the conformations into 5 or 10 clusters using K-means clustering [22, 23], selecting the cluster with the lowest mean RMSD as the folded state. From this cluster, native contacts were defined as those which were closer than 10 Å for more than 80% of the time, only considering C_α atoms, and only atom-pairs separated by more than four (for the shorter systems: GB1-hairpin, Trp-cage and Beta3s) or seven (for the larger Top7-Cfr) residues along the sequence. Using these contacts, the Q-factor was then defined as

$$Q = \sum_{i=1}^{N_{\text{aa}}} \sum_{j=1}^{n_i} \frac{1}{1 + e^{10(d_{ij} - (d_{ij}^0 + 1))}} \Bigg/ \sum_{i=1}^{N_{\text{aa}}} n_i \qquad [2]$$

where N_{aa} is the number of amino acid resiudes, n_i is the number of contacts of residue i, d_{ij} is the C_α-C_α distance between residue i and residue j, and d_{ij}^0 is the distance between the same contacts in the native state. According to the Q value, the trajectory is partitioned into segments: folded($Q > 0.9$), unfolded ($Q < 0.1$) and "transition path" segments. At any given point in time along the trajectory, the system is labeled as an up walker (+) or a down walker (−). The label '+' is left unchanged upon visits to $Q = 0.1$ but changed to '−' when it reaches $Q = 0.9$. Let τ_{up} and τ_{down} represent the average length of the up-walker and down-walker trajectory segment, respectively. The round-trip time is counted as the time (MC steps) the system takes to move from one boundary to the other and back again, which can be found as $\tau = \tau_{\text{up}} + \tau_{\text{down}}$. The round-trip time of all our simulations are shown in Fig. 3 and the time evolution of the Q-factor reaction coordinate is illustrated in Fig. S5.

Performance The CS-TORUS represents no significant performance bottleneck when used in simulation. As is normally the case, the pairwise interactions from the force field dominate the computational cost. The simulations in this study were done using the PROFASI force field, which is extremely efficient due to an efficient caching mechanism, and the choice of rather short pair-wise cutoffs [24]. Even with this choice of force-field, however, there was no significant impact when using the CS-TORUS model compared to the unbiased case.

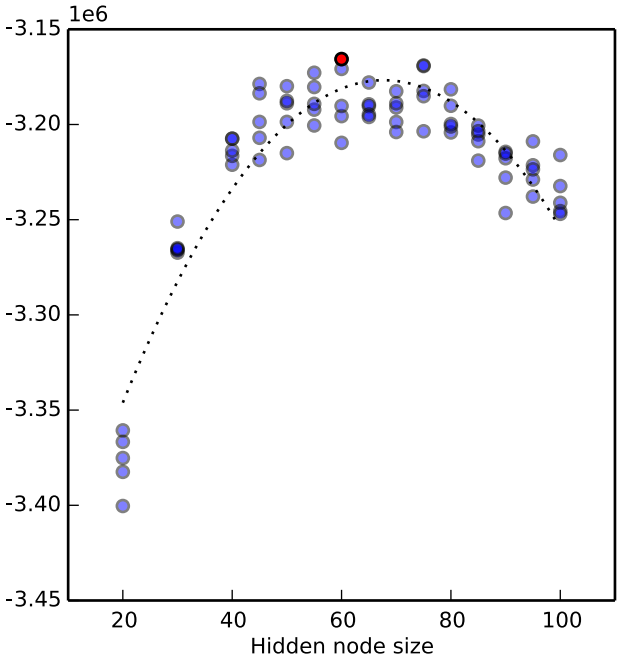

Fig. S1. Bayesian Information Criterion (BIC) [25] scores for different values of the hidden node size. For each size, five models were trained. The red circle represents the model chosen for the simulations in this article. For the Ubiquitin simulations in the article, a separate model was used (since these simulations were added later, and Ubiquitin was not excluded from the training set for the original model). This model was trained using the exact same procedure, on a data set excluding any Ubiquitin related proteins (see main text for the selection criteria), resulting in a similar BIC curve.

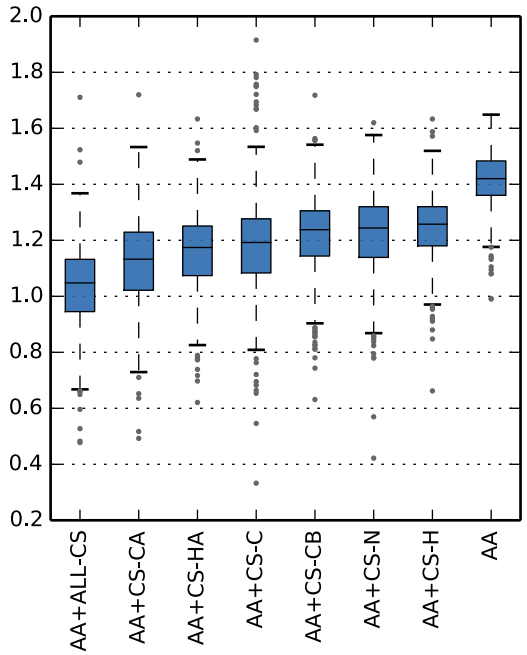

Fig. S2. Comparison of sampling accuracy of the model when using different chemical shift atom types (CA, CB, C, H, HA, N), in addition to the amino acid information (AA). The plot shows the average angular deviation (in radians) from the crystal structure for all proteins in the training set for which chemical shifts for all six atom types were recorded. For each protein, 10 samples were drawn from the model, and the angular deviation was calculated as described in ref. [6].

Table S1. The coverage of chemical shift data for each of the simulated proteins. The numbers specify the fraction of residues for which a given chemical shift value was available. N/A is used to highlight that a given type of chemical shift value were not available for any of the residues.

protein	length	C	CA	CB	HA	H	N
GB3	56	0.96	1.00	0.93	0.98	0.93	0.98
Ubiquitin	76	0.91	1.00	0.91	0.95	0.92	0.92
GB1-hairpin	16	1.00	1.00	0.94	N/A	N/A	1.00
Trp-cage	20	N/A	N/A	N/A	0.75	0.85	N/A
Beta3s	20	N/A	1.00	N/A	0.95	1.00	N/A
Top7-Cfr	49	0.96	0.86	0.84	0.98	0.94	0.94

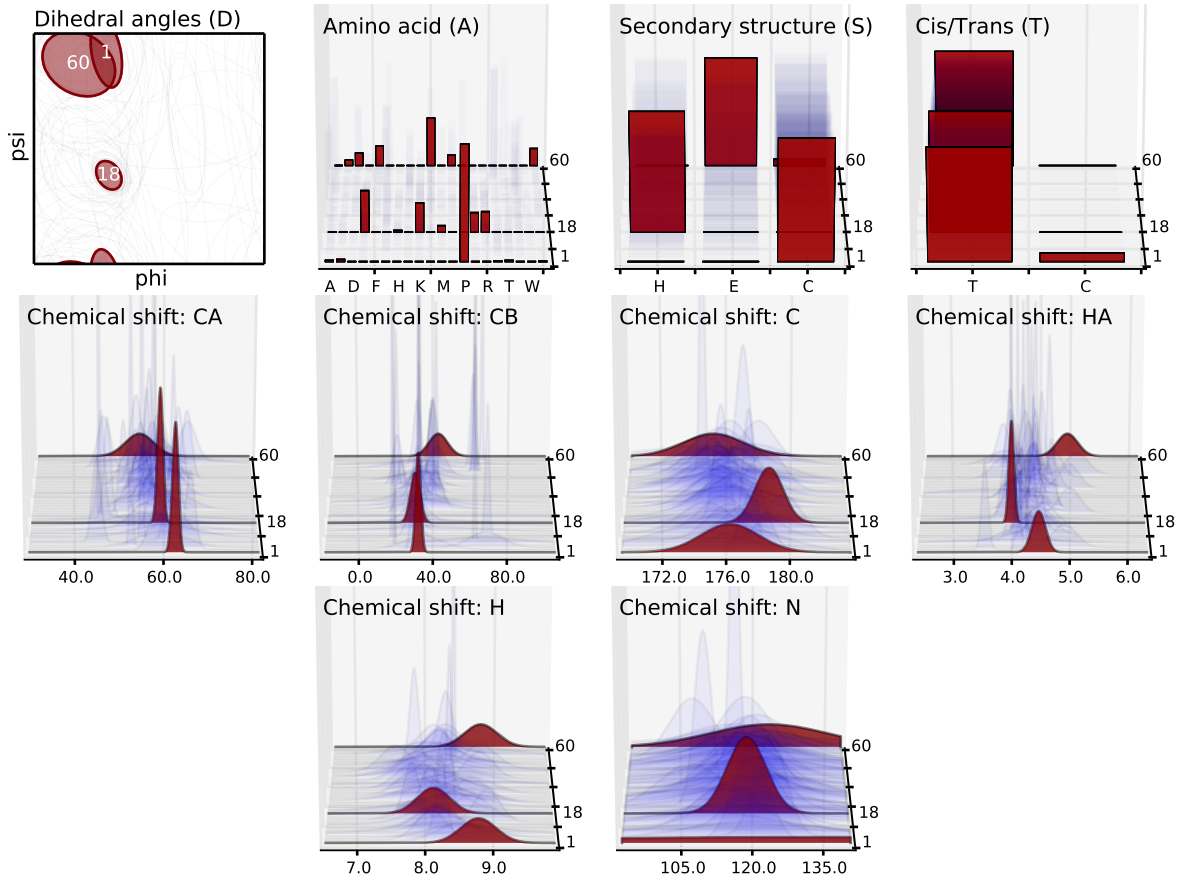

Fig. S3. Examples of emission probabilities for three hidden node states, representing three different secondary structure preferences: state 18 is helix prone, state 60 corresponds to beta structures while state 1 is mainly used for Prolines, and is one of the few states with a non-zero probability for adopting a cis state. Note that since Proline has no H (HN) atom, the H distribution for node 1 is due to the other (sparsely populated) amino acids for this node. Similarly, the N signal for node 1 is extremely broad, which is presumably due to the fact that Proline N atoms are often not assigned in NMR experiments since they do not have an associated hydrogen (and are therefore not visible in HSQC-based experiments).

Original training set: 108m, 109m, 1a43, 1a5j, 1a6j, 1a7g, 1a91, 1aab, 1ab1, 1aba, 1adw, 1ae3, 1aep, 1af8, 1agt, 1ahl, 1ail, 1ake, 1akh, 1akp, 1ans, 1aoo, 1apo, 1aq5, 1aq5, 1ass, 1atb, 1atx, 1auu, 1avs, 1axh, 1az6, 1b0c, 1b10, 1b1v, 1b2v, 1b3c, 1b56, 1b5a, 1b64, 1b72, 1b88, 1bbi, 1bbl, 1bci, 1bcx, 1bd9, 1bdo, 1bds, 1bed, 1bf4, 1bfc, 1bgf, 1bhi, 1bhu, 1bi7, 1bja, 1bk8, 1bku, 1bm4, 1bm9, 1bnz, 1bo0, 1bpv, 1bqv, 1bqz, 1bri, 1brj, 1bv8, 1bw5, 1bwo, 1bwx, 1bxl, 1byf, 1bzb, 1c3t, 1c49, 1c4z, 1c55, 1c5a, 1c6w, 1c76, 1c7f, 1c7w, 1c89, 1c8a, 1c8c, 1cb9, 1cbh, 1ccv, 1ce3, 1cex, 1cfe, 1cho, 1cix, 1cku, 1ckv, 1ckw, 1ckx, 1cm2, 1cmz, 1cnr, 1col, 1com, 1com, 1cpz, 1cqm, 1cr8, 1crb, 1csg, 1cw5, 1cw6, 1cx1, 1cxw, 1cy5, 1cz5, 1d03, 1d1d, 1d1o, 1d5g, 1d8k, 1d9s, 1dav, 1dbd, 1dcd, 1dcj, 1dd2, 1dd5, 1de3, 1df6, 1dfj, 1dfu, 1dhn, 1div, 1dk0, 1dk3, 1dkc, 1dl0, 1dp3, 1dpu, 1dqc, 1dqe, 1dsb, 1dtk, 1du9, 1dv0, 1dwy, 1dx7, 1dx8, 1dyt, 1e0e, 1e0m, 1e17, 1e3y, 1e8b, 1e9t, 1edn, 1egj, 1egx, 1eh1, 1ehx, 1eih, 1eij, 1eik, 1ejf, 1ejm, 1ejq, 1ejq, 1ek8, 1el0, 1emx, 1emz, 1enf, 1eoq, 1epg, 1et1, 1ev0, 1exk, 1exp, 1eyf, 1ez9, 1ezg, 1ezt, 1f0z, 1f2l, 1f2m, 1f53, 1f62, 1f81, 1f8h, 1f94, 1f95, 1fbr, 1fd3, 1fd9, 1fdq, 1fe4, 1fex, 1ffj, 1fho, 1fil, 1fj7, 1fjc, 1fkh, 1fmm, 1fo1, 1fov, 1fpw, 1ftt, 1fu9, 1fwo, 1fzt, 1fzy, 1g03, 1g26, 1g2h, 1g47, 1g4c, 1g4f, 1g5v, 1g6a, 1g6h, 1g6m, 1g6p, 1g7f, 1g7o, 1g8i, 1g9e, 1g9p, 1gaw, 1ggq, 1ggw, 1gh9, 1gjx, 1gk5, 1gl5, 1gn0, 1gnu, 1go5, 1gpr, 1guj, 1gwy, 1gxe, 1gxq, 1h0j, 1h0z, 1h20, 1h2o, 1h3z, 1h4a, 1h4b, 1h67, 1h70, 1h7y, 1h8b, 1ha6, 1ha8, 1ha9, 1hb8, 1hc9, 1hc9, 1hcb, 1hcc, 1hd6, 1hej, 1hfc, 1hfz, 1hh8, 1hhn, 1hj0, 1hll, 1hoe, 1hof, 1hp2, 1hpc, 1hq2, 1hqb, 1hs5, 1hst, 1hum, 1huu, 1hvw, 1hzk, 1i11, 1i1j, 1i25, 1i26, 1i2v, 1i4f, 1i5k, 1i6f, 1i8x, 1iar, 1iaz, 1ibi, 1ica, 1icf, 1ich, 1ie5, 1ieh, 1ifc, 1ifw, 1igv, 1iho, 1ihq, 1ijp, 1ijz, 1iko, 1ilo, 1ip0, 1ip2, 1ipb, 1iqs, 1irr, 1irz, 1iu1, 1iv6, 1iv7, 1ivm, 1ivo, 1iw0, 1iw4, 1iwt, 1ix5, 1iyc, 1iym, 1iyr, 1iyt, 1j0f, 1j0t, 1j26, 1j2n, 1j3g, 1j3t, 1j54, 1j56, 1j5j, 1j5k, 1j7d, 1j7h, 1j7m, 1j8k, 1j9i, 1jaj, 1jas, 1jba, 1jbi, 1jbj, 1jc2, 1jc6, 1jcu, 1jdq, 1je3, 1jfj, 1jfn, 1jgk, 1jh3, 1jiw, 1jjd, 1jjg, 1jjz, 1jkn, 1jl9, 1jlz, 1jns, 1jo6, 1joc, 1jr2, 1jr6, 1jrm, 1js2, 1jse, 1ju8, 1jvo, 1jw2, 1jw3, 1jwe, 1jxc, 1jzu, 1k36, 1k42, 1k46, 1k7b, 1ka5, 1kba, 1kcy, 1kdj, 1kf3, 1kf8, 1kft, 1kg1, 1kio, 1kj6, 1kjt, 1kkg, 1kma, 1kmv, 1kmx, 1kn6, 1kn7, 1kqq, 1kri, 1krw, 1ks0, 1ktm, 1ktz, 1kv4, 1kvn, 1kvz, 1kx9, 1kzw, 1l1m, 1l1p, 1l2m, 1l3g, 1l3y, 1l4s, 1l4v, 1l6t, 1l6x, 1l7b, 1l7y, 1lbj, 1ld6, 1lfc, 1liq, 1lkj, 1ll8, 1lmm, 1lo1, 1lp1, 1lsi, 1lu0, 1lw6, 1lwm, 1lwr, 1m25, 1m2f, 1m31, 1m36, 1m3g, 1m3v, 1m4f, 1m58, 1m5z, 1m7t, 1m8a, 1m9l, 1mb6, 1mft, 1mg8, 1mhi, 1mi4, 1mid, 1mjc, 1mke, 1mm2, 1mm3, 1mmq, 1mms, 1mox, 1mpz, 1mr6, 1msz, 1mtx, 1mwz, 1mxe, 1myo, 1n0s, 1n0z, 1n27, 1n2d, 1n4c, 1n6z, 1n8l, 1n9d, 1n9j, 1naq, 1nbl, 1nd4, 1ne3, 1ne5, 1nee, 1nei, 1ner, 1ng2, 1ng7, 1nho, 1ni7, 1njq, 1nla, 1nlp, 1nmr, 1nnv, 1no8, 1noa, 1nq4, 1nr3, 1nra, 1nso, 1ntc, 1nw2, 1nw2, 1nxi, 1ny4, 1ny8, 1ny9, 1nya, 1nzp, 1o5u, 1o6x, 1o82, 1o8r, 1o8t, 1ob1, 1obo, 1oc0, 1off, 1oh1, 1okh, 1om2, 1omy, 1onj, 1op4, 1oqk, 1oqr, 1orl, 1os3, 1os3, 1osp, 1ov2, 1ovh, 1ovx, 1ow5, 1owa, 1owx, 1oy2, 1ozz, 1p2o, 1p4w, 1p68, 1p6q, 1p6r, 1p6s, 1p6t, 1p6z, 1p7a, 1p7m, 1p8a, 1p8g, 1p94, 1p9j, 1pa4, 1pb5, 1pba, 1pd7, 1pex, 1pey, 1pf9, 1pfj, 1php, 1pht, 1pjz, 1pn5, 1poh, 1ppf, 1ps2, 1pt4, 1pu1, 1pun, 1puz, 1pv0, 1pv3, 1pve, 1pvz, 1pz4, 1q02, 1q2f, 1q2u, 1q2z, 1q56, 1q5f, 1q6u, 1q71, 1q7x, 1q8k, 1q8l, 1q8x, 1qa9, 1qav, 1qe6, 1qfn, 1qg7, 1qj8, 1qk7, 1qk9, 1qky, 1qlo, 1qlx, 1qlz, 1qm3, 1qog, 1qry, 1qst, 1qu6, 1qwv, 1qxf, 1qxn, 1qz8, 1r1t, 1r1u, 1r21, 1r3b, 1r4k, 1r5e, 1r5r, 1r5s, 1r69, 1r6h, 1r6r, 1r73, 1r79, 1r7j, 1r8u, 1r9k, 1r9p, 1rdg, 1rdu, 1rg6, 1rgh, 1rgw, 1ri0, 1rji, 1rjj, 1rjv, 1rkl, 1rl1, 1rmk, 1rn1, 1rpg, 1rpj, 1rq6, 1rq8, 1rqu, 1rro, 1rrz, 1rw2, 1rwu, 1rwy, 1ry4, 1ryk, 1rzs, 1rzw, 1s04, 1s3s, 1s40, 1s62, 1s6d, 1s6i, 1s6j, 1s6l, 1s6n, 1s6u, 1sa8, 1sai, 1sb6, 1sce, 1sce, 1se9, 1sf0, 1sfc, 1sg7, 1sh1, 1siy, 1sj6, 1sjq, 1sjr, 1sko, 1sm7, 1snc, 1snl, 1snm, 1sou, 1sq8, 1sr2, 1srb, 1srk, 1srz, 1ss6, 1ssl, 1st7, 1sxd, 1sxe, 1sxl, 1t0g, 1t0k, 1t0y, 1t17, 1t1h, 1t2y, 1t3k, 1t4z, 1tba, 1tcf, 1te4, 1te7, 1th5, 1ti3, 1tiz, 1tjf, 1tkv, 1tn3, 1top, 1tot, 1tp9, 1tph, 1tph, 1tpk, 1tpx, 1tq1, 1tqz, 1tte, 1ttg, 1ttx, 1tuk, 1tuz, 1tvg, 1tvi, 1tvj, 1tvq, 1tw4, 1two, 1txe, 1txx, 1tym, 1u06, 1u07, 1u2g, 1u2p, 1u3m, 1u5l, 1u5m, 1u5s, 1u6f, 1u7j, 1u89, 1uap, 1ub1, 1ubq, 1uc6, 1uck, 1ud7, 1udr, 1ue9, 1uem, 1ueo, 1uep, 1uew, 1uff, 1ufg, 1ufm, 1ufn, 1ufx, 1ug7, 1ugl, 1uhf, 1uhi, 1uht, 1uhu, 1ujo, 1ujs, 1ujt, 1uju, 1ujv, 1ujx, 1ukx, 1ul7, 1umq, 1uoh, 1utx, 1uuc, 1uug, 1uw0, 1uzc, 1v31, 1v32, 1v4r, 1v5k, 1v5m, 1v5n, 1v5p, 1v5q, 1v5r, 1v5s, 1v5u, 1v63, 1v66, 1v6p, 1v6r, 1v86, 1v88, 1v9v, 1v9w, 1va9, 1vae, 1vb0, 1vc1, 1vcx, 1vd0, 1vdi, 1vdq, 1vj6, 1vp6, 1vpc, 1vsa, 1vyf, 1vyn, 1w0t, 1w41, 1w6b, 1w6v, 1w80, 1wcj, 1wej, 1wey, 1wez, 1wf1, 1wf2, 1wf5, 1wf9, 1wfg, 1wfi, 1wfj, 1wfm, 1wfn, 1wfo, 1wfq, 1wfs, 1wft, 1wfv, 1wfw, 1wfy, 1wfz, 1wg5, 1wgq, 1wgr, 1wgs, 1wgu, 1wgv, 1wgw, 1wgy, 1wh3, 1wh4, 1wh5, 1wh6, 1wh7, 1wh8, 1wh9, 1wha, 1whn, 1whr, 1whu, 1wi0, 1wi8, 1wik, 1wil, 1win, 1wj1, 1wjd, 1wji, 1wjj, 1wjk, 1wjl, 1wjn, 1wjo, 1wjp, 1wjq, 1wjr, 1wjs, 1wjt, 1wju, 1wjz, 1wk0, 1wk1, 1wkt, 1wkx, 1wlm, 1wlx, 1wpi, 1wqk, 1wqq, 1wqu, 1wt7, 1wtq, 1wu0, 1wum, 1wvk, 1wwy, 1wxl, 1wxn, 1wyl, 1wyn, 1wyo, 1wyw, 1wzv, 1x05, 1x1f, 1x1g, 1x22, 1x32, 1x3q, 1x5b, 1x5i, 1x5k, 1x6b, 1x6d, 1x6e, 1x6f, 1x6h, 1x8r, 1x9a, 1x9b, 1xbl, 1xd3, 1xdg, 1xfl, 1xhj, 1xhs, 1xjh, 1xke, 1xlq, 1xm0, 1xmt, 1xn5, 1xn6, 1xn7, 1xn9, 1xna, 1xne, 1xo8, 1xoy, 1xpa, 1xpn, 1xq8, 1xrd, 1xrk, 1xs8, 1xsc, 1xsf, 1xsw, 1xu6, 1xwe, 1xwn, 1xyj, 1xyk, 1xyq, 1xyw, 1y0j, 1y0j, 1y15, 1y1b, 1y1c, 1y2g, 1y4o, 1y5k, 1y62, 1y7n, 1y93, 1ycq, 1ydu, 1yel, 1yez, 1yh5, 1yjt, 1ykg, 1yky, 1yla, 1ynr, 1yob, 1yp7, 1yqa, 1ysb, 1ysm, 1yu7, 1yua, 1yvc, 1yws, 1ywu, 1yww, 1yx3, 1yyb, 1yzc, 1z1m, 1z3r, 1z6h, 1z6s, 1z9i, 1zdn, 1zdv, 1zfs, 1zgu, 1zit, 1zk6, 1zkh, 1zli, 1zlq, 1znd, 1zq3, 1zr7, 1zr9, 1zrf, 1zts, 1zu1, 1zu2, 1zv6, 1zwv, 1zyn, 1zza, 1zzp, 2a00, 2a0a, 2a0b, 2a0n, 2a2p, 2a36, 2a37, 2a4h, 2a5d, 2a5e, 2a7y, 2adf, 2adr, 2adz, 2afg, 2afj, 2afp, 2aih, 2aje, 2akk, 2akl, 2al3, 2al4, 2alg, 2aoj, 2aq0, 2arw, 2asw, 2asy, 2av5, 2avg, 2axd, 2axl, 2ayj, 2ayx, 2b3a, 2b3i, 2b3w, 2b59, 2b5x, 2b6f, 2b7e, 2b86, 2b88, 2b89, 2b8x, 2b95, 2bay, 2bbg, 2bc5, 2bem, 2bf5, 2bid, 2bjx, 2bky, 2bky, 2bl5, 2bru, 2buo, 2bvo, 2bvu, 2bye, 2bz2, 2bzb, 2c0s, 2c6y, 2ca5, 2cbs, 2cdn, 2cg7, 2ch4, 2cjr, 2cly, 2cnj, 2cnp, 2cnr, 2co8, 2coa, 2coc, 2cod, 2cof, 2com, 2cpb, 2cu7, 2cue, 2cuf, 2cum, 2cwi, 2d07, 2d3g, 2d7m, 2d7n, 2d7o, 2d7p, 2d7q, 2d82, 2d9t, 2d9y, 2d9z, 2dbj, 2dc2, 2dez, 2dgc, 2dhj, 2di8, 2di9, 2dia, 2dib, 2dic, 2diz, 2dj4, 2djs, 2dk9, 2dkq, 2dlg, 2dmb, 2dmc, 2dml, 2dmq, 2dn6, 2dn7, 2do8, 2dtq, 2e29, 2e45, 2e6i, 2eb8, 2ecc, 2ech, 2end, 2ers, 2esp, 2etl, 2ewl, 2exd, 2exf, 2exn, 2ezh, 2f05, 2f09, 2f1e, 2f30, 2f3y, 2f3z, 2f5m, 2f91, 2fa4, 2fb7, 2fe0, 2ffk, 2fft, 2fi2, 2fj3, 2fj6, 2fjy, 2fk4, 2fke, 2fki, 2fm4, 2fnb, 2fnf, 2frw, 2fs1, 2fvn, 2fvt, 2fxp, 2fy9, 2fz5, 2g0l, 2g0u, 2g1d, 2g7j, 2g9j, 2g9j, 2g9o, 2ga5, 2gab, 2gbs, 2ge9, 2git, 2gjf, 2gl1, 2gm2, 2gmg, 2goo, 2gov, 2gqb, 2gs0, 2gtg, 2gvs, 2gw6, 2gyk, 2gyt, 2gzo, 2gzz, 2h0p, 2h2r, 2h3j, 2h3k, 2h5m, 2h7a, 2h80, 2hcc, 2hdm, 2heq, 2hg7, 2hga, 2hgc, 2hgk, 2hgu, 2hh8, 2hi3, 2hi6, 2hj8, 2hjj, 2hjq, 2hoa, 2hpu, 2hsh, 2hst, 2hsx, 2htj, 2hym, 2hym, 2i32, 2i3b, 2i9a, 2i9y, 2ida, 2ido, 2idy, 2if1, 2ife, 2iim, 2iln, 2ilx, 2in2, 2io2, 2ion, 2irf, 2itl, 2j03, 2j4t, 2jm4, 2jmp, 2jmu, 2jn0, 2jn4, 2jn6, 2jn7, 2jn9, 2jna, 2jne, 2jng, 2jnu, 2jny, 2jo6, 2joe, 2jol, 2joq, 2jov, 2joy, 2joz, 2jq5, 2jqo, 2jr5, 2jra, 2jrj, 2jrm, 2jrp, 2jrr, 2jrz, 2js4, 2ktx, 2mb5, 2mfn, 2mob, 2nmo, 2noc, 2npr, 2nwg, 2nwm, 2nwt, 2nxn, 2o3b, 2oa4, 2oi3, 2orc, 2out, 2ovo, 2ow9, 2pea, 2pjf, 2pjg, 2pji, 2pkt, 2plh, 2pp4, 2pph, 2pst, 2q00, 2rn2, 2sgd, 2sni, 2uub, 2uyz, 2uzg, 2vpf, 2xbd, 2z2i, 3eza, 3icb, 3lri, 3ncm, 3pdz, 3pyp, 3ssi, 3wrp, 451c, 4ake, 4hir, 4icb, 5cro, 5hpg, 5pnt, 7hsc, 7rxn, 8abp, 8tfv, 9pcy

Ubiquitin training set: excluded compared to above 1c3t, 1gnu, 1kjt, 1l7y, 1mg8, 1s3s, 1se9, 1t0y, 1ubq, 1ud7, 1v86, 1wf9, 1wfy, 1wgr, 1wgy, 1wh3, 1wjn, 1wju, 1wyw, 1xd3, 1zgu, 1zkh, 2al3, 2b3a, 2bye, 2d07, 2d3g, 2hj8, 2io2, 2pea, 2uyz

Ubiquitin training set: included compared to above 1b1h, 1clv, 1cwc, 1cyn, 1eio, 1f2r, 1f3v, 1f93, 1fcl, 1fd6, 1h4h, 1i8h, 1lq7, 1m15, 1mw4, 1q10, 1rb9, 1rx2, 1rx4, 1skm, 1u7e, 1uea, 1uwx, 1x27, 1xct, 1zxh, 2aiz, 2axi, 2brz, 2c7p, 2crd, 2fi4, 2fi5, 2g46, 2gfe, 2h61, 2h61, 2hze, 2pg1, 2pg1, 2pld, 2psp

Fig. S4. PDB IDs of proteins used as the training set of the model. The Ref-DB [1] chemical shift annotated files are available for download from `http://refdb.wishartlab.com/`.

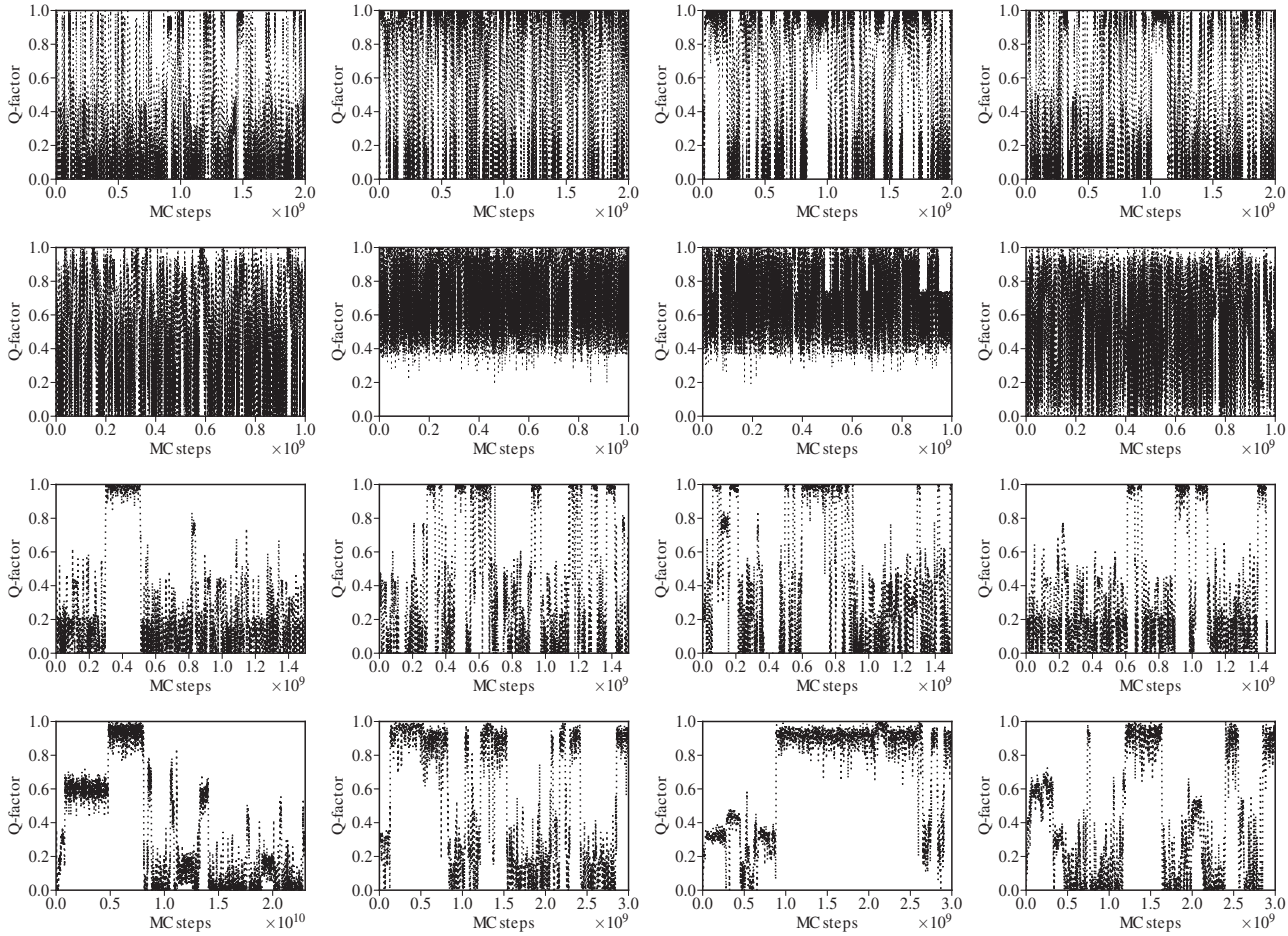

Fig. S5. Run time trajectories from a representative thread from each simulation. From the left to the right column: Unbiased simulation, Biased move simulation, Biased energy simulation, Compensated bias simulation. From the top to the bottom row: GB1-hairpin, Trp-cage, Beta3s, Top7-Cfr.

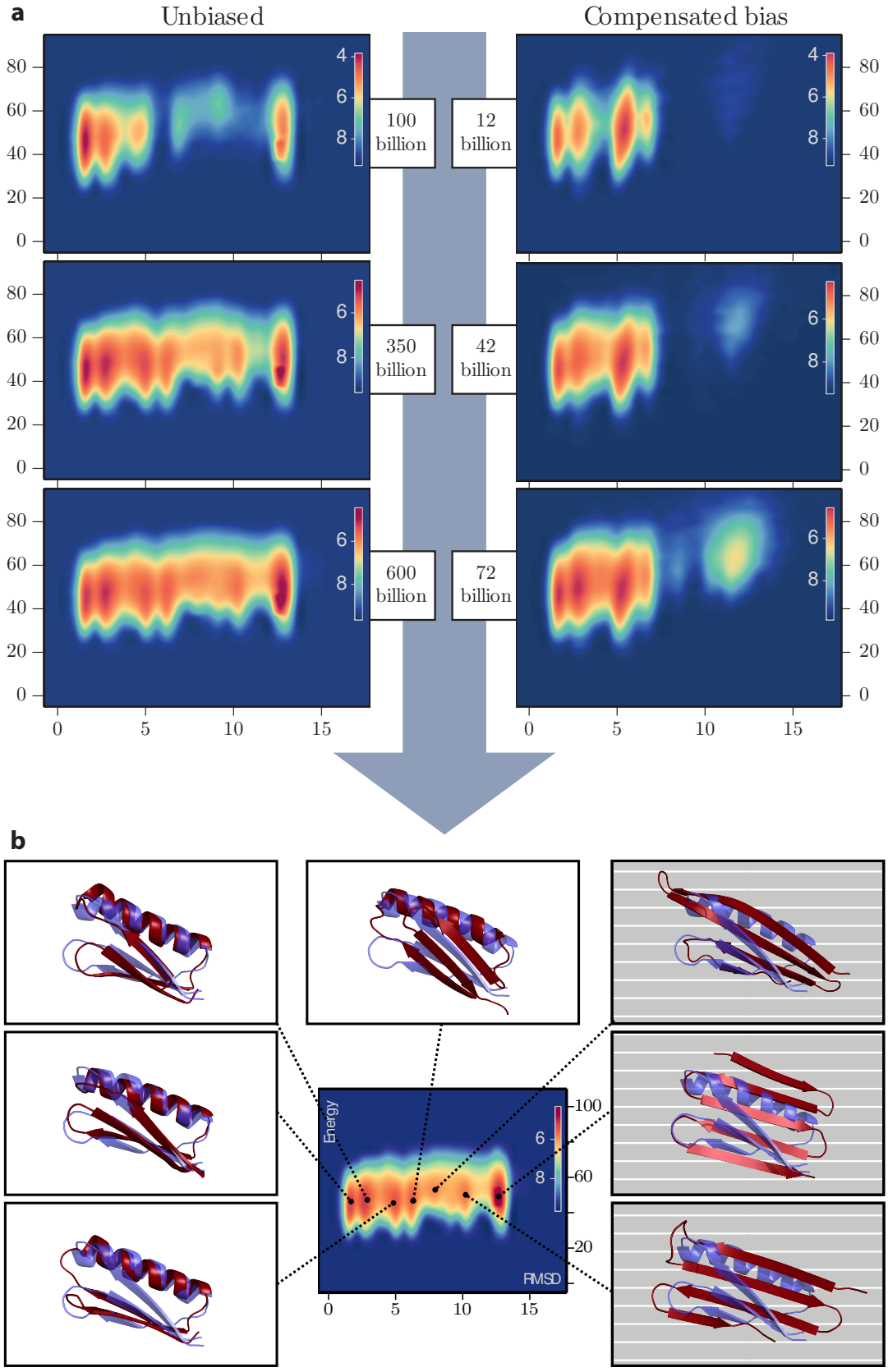

Fig. S6. Analysis of the unbiased and the compensated bias simulations for the Top7-Cfr system. a) The progression of convergence over time (number of iterations), b) The structures corresponding to the different peaks in the free energy landscape (native structure in blue). The structures corresponding to the high RMSD peaks (in grey) have beta structure instead of the helix, which is at odds with the chemical shift signal, and is therefore never sampled by CS-TORUS.

Table S2. Five strategies for simulations using a generative probabilistic model. 1) Standard Metropolis-Hastings simulation with uniform proposal distribution. 2) Biased simulation using biased moves. 3) Biased simulation using an explicit bias in the potential. 4) The standard correction when conducting Metropolis-Hastings simulations with a non-uniform sampling bias. 5) Compensated bias simulations, using generalized ensembles a range of γ values are explored; at $\gamma = 1$, this corresponds to method 4). Variants 1, 2, 3, and 5 were explored in this paper (see Fig. 3(a), 3(b), 3(c), and 3(d), respectively).

	Proposal	Evaluation in acceptance criterion	Effective probability target
1	uniform	$\exp(-\beta E(x))$	$\exp(-\beta E(x))$
2	$P_{DBN}(x)$	$\exp(-\beta E(x))$	$\exp(-\beta E(x))P_{DBN}(x)$
3	uniform	$\exp(-\beta E(x) + \ln(P_{DBN}(x)))$	$\exp(-\beta E(x))P_{DBN}(x)$
4	$P_{DBN}(x)$	$\exp(-\beta E(x) - \ln(P_{DBN}(x)))$	$\exp(-\beta E(x))$
5	$P_{DBN}(x)$	$\exp(\gamma(-\beta E(x) - \ln(P_{DBN}(x))))$	$\gamma = 0 \Rightarrow P_{DBN}$ $\gamma = 1 \Rightarrow \exp(-\beta E(x))$

References

1. Zhang H, Neal S, Wishart DS (2003) Refdb: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25: 173–195.
2. Murzin AG, Brenner SE, 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.
3. Sillitoe I et al. (2013) New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures. Nucleic acids res 41: D490–D498.
4. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
5. Nielsen SF (2000) The stochastic EM algorithm: estimation and asymptotic results. Bernoulli 6: 457–89.
6. Boomsma W et al. (2008) A generative, probabilistic model of local protein structure. Proc Natl Acad Sci USA 105: 8932–8937.
7. Cawley SL, Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. Bioinformatics 19 Suppl 2: ii36–41.
8. Hamelryck T, Kent JT, Krogh A (2006) Sampling realistic protein conformations using local structural bias. PLoS Comput Biol 2: e131.
9. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.
10. Zimm BH, Bragg J (1959) Theory of the phase transition between helix and random coil in polypeptide chains. J Chem Phys 31: 526–535.
11. Grosberg AY, Khokhlov AR (1994) Statistical physics of macromolecules (American Institute of Physics New York).
12. Irbäck A, Mohanty S (2006) PROFASI: a monte carlo simulation package for protein folding and aggregation. J Comput Chem 27: 1548–1555.
13. Favrin G, Irbäck A, Sjunnesson F (2001) Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. J Chem Phys 114: 8154–8158.
14. Ferkinghoff-Borg J (2002) Optimized Monte Carlo analysis for generalized ensembles. Eur. Phys. J. B 29: 481–484.
15. Frellsen J (2011) Ph.D. thesis (University of Copenhagen) http://muninn.sourceforge.net/.
16. Ferkinghoff-Borg J (2012) Monte carlo methods for inference in high-dimensional systems. Bayesian Methods in Structural Bioinformatics, Statistics for Biology and Health, eds Hamelryck T, Mardia K, Ferkinghoff-Borg J (Springer), pp 49–93.
17. Boomsma W, Frellsen J, Hamelryck T (2012) Probabilistic models of local biomolecular structure and their applications. Bayesian Methods in Structural Bioinformatics, Statistics for Biology and Health, eds Hamelryck T, Mardia K, Ferkinghoff-Borg J (Springer), pp 233–254.
18. Best RB, Hummer G (2010) Coordinate-dependent diffusion in protein folding. Proc Natl Acad Sci USA 107: 1088–1093.
19. Best RB, Hummer G (2005) Reaction coordinates and rates from transition paths. Proc Natl Acad Sci USA 102: 6732–6737.
20. Best RB, Hummer G, Eaton WA (2013) Native contacts determine protein folding mechanisms in atomistic simulations. Proc Natl Acad Sci USA 110: 17874–17879.
21. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. Science 334: 517–520.
22. MacQueen J (1967) 5th Berkeley symposium on mathematical statistics and probability. Berkeley, CA.
23. Harder T, Borg M, Boomsma W, Røgen P, Hamelryck T (2012) Fast large-scale clustering of protein structures using Gauss integrals. Bioinformatics 28: 510–515.
24. Irbäck A, Mitternacht S, Mohanty S (2009) An effective all-atom potential for proteins. PMC Biophysics 2: 2.
25. Schwarz G, et al. (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.