

Lecture 11

Nuclear Vector Replacement

This lecture presents the NVR algorithm, and its applications in automated NMR resonance assignments [1, 2] and 3D structure homology detection [3, 4].

1 Experimental Input

The following data are processed in the NVR algorithm: unassigned chemical shifts in H^N - ^{15}N HSQC spectra, H^N - ^{15}N RDCs in two media, amide exchange data, and unassigned peaks in 3D ^{15}N -NOESY spectra. The amide exchange information provides probabilistic classifications for the peaks in the HSQC corresponding to non hydrogen-bonded, solvent accessible backbone amide protons. All the H^N - ^{15}N RDC, H-D exchange and ^{15}N -NOESY serve as the geometric constraints on assignment. Fig 1 shows the information contents and roles of the input data.

2 Nuclear Vector Replacement

The NVR algorithm consists of three phases: *Tensor Estimation*, *Resonance Assignment*, and *Structure Refinement*, as shown in Figure 2. In the first phase, NVR estimates the alignment tensors in both media. In the second phase, NVR applies the estimated tensors, and uses an iterative computational process, such as a Bayesian or Expectation/Maximization

Experiment/Data	Information Content	Role
H^N - ^{15}N HSQC	H^N , ^{15}N Chemical shifts	Backbone resonances, Cross-referencing NOESY
H^N - ^{15}N RDC (in 2 media)	Restrains on amide bond vector orientation	Tensor Determination, Resonance Assignment,
H-D exchange HSQC	Identifies solvent exposed amide protons	Tensor Determination
H^N - ^{15}N HSQC-NOESY	Distance restraints between spin systems	Tensor Determination, Resonance Assignment
^{15}N TOCSY	Side-Chain Chemical Shifts	Tensor Determination, Resonance Assignment
Backbone Structure	Tertiary Structure	Tensor Determination, Resonance Assignment
Chemical Shift Predictions	Restrains on Assignment	Tensor Determination, Resonance Assignment

Figure 1: Information contents and roles of the input data [4].

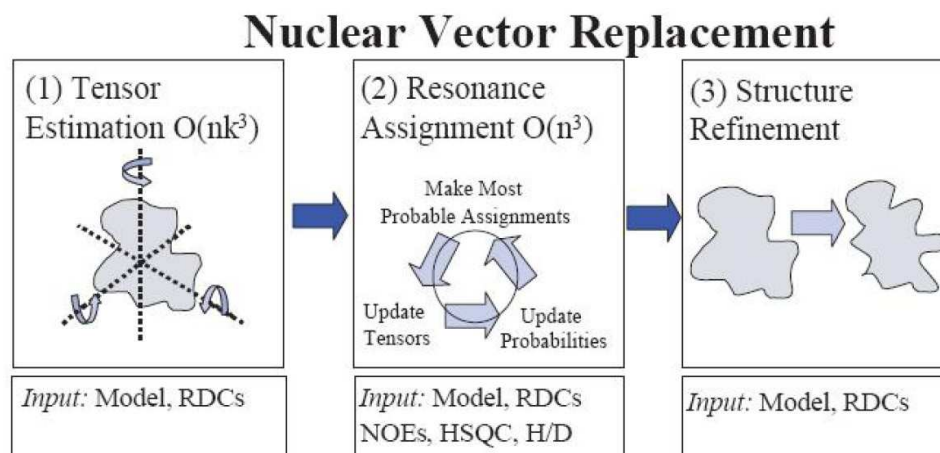


Figure 2: Nuclear Vector Replacement [2].

framework, to search the optimal resonance assignments. After resonance assignments, the assigned RDCs are used to refine the structure of the model by using the algorithm in [6]. In the following subsections, we will present more details on the tensor estimation and resonance assignment phases.

2.1 Tensor Estimation

An alignment tensor S is a symmetric and traceless 3×3 matrix with five degrees of freedom, i.e., three Euler angles α , β and γ , and the axial D_a and rhombic D_r components of an ellipsoid scaling the dipolar couplings. The alignment tensor can be diagonalized as follows:

$$S = V\Sigma V^T$$

where $V \in SO(3)$ is a 3×3 rotation matrix that defines the *principle order frame*, and Σ is a 3×3 diagonal and traceless matrix containing the eigenvalues of S and encoding D_a and D_r .

The diagonal elements of Σ can be obtained based on the powder pattern method. Given Σ , a rotation matrix $V(\alpha, \beta, \gamma)$, corresponding to an alignment tensor S , is found such that the Kullback-Leibler distance between the distribution of unassigned experimental RDCs and the distribution of the back-computed RDCs.

Time Complexity. The eigenvalues of S are computed in $O(nk^3)$ time, where n is the number of experimental RDCs and k is the resolution of the search-grid over $SO(3)$. The rotation matrix is calculated in $O(nk^3)$ time. Given Σ and V , constructing S takes only $O(1)$ time. Thus, the total running time for tensor estimation is $O(nk^3)$.

2.2 Resonance Assignment

Based on the estimated tensors S_1 and S_2 in two media, the back-computed RDCs are obtained according to the following equation:

$$D = D_{\max} v^T S_i v, \quad i = 1, 2,$$

where D_{\max} is the dipolar interaction constant, and v is the internuclear vector orientation.

Let Q be the set of HSQC peaks, R be the set of residues, D_m be the set of experimental RDCs in the medium m , and B_m be the back-computed RDCs in the medium m , based on S_m . Let M_m be the assignment probability matrix for the medium m such that:

$$M_m(q, r) = \Pr(q \mapsto r | S_m) = N(d_m(q) - b_m(r, S_m), \mu_m, \sigma_m), \quad (1)$$

where $q \in Q$, $r \in R$, $d_m(q) \in D_m$, $b_m(r, S_m) \in B_m$, and $N(d_m(q) - b_m(r, S_m), \mu_m, \sigma_m)$ denotes the Normal distribution function with the observing variable $d_m(q) - b_m(r, S_m)$, the mean μ_m , and the standard deviation σ_m . Each entry in M_m will be set to zero if the assignment $q \mapsto r$ violates some geometric constraints, such as the d_{NNs} from ^{15}N -NOESY spectra, or amide exchange. The d_{NNs} provides distance constraints between peaks that can correlated to the residues in the model, while the H-D exchange information identifies fast exchanging amide protons that are likely to be solvent-exposed and non-hydrogen bounded, and can be correlated to the structure model.

Equation (1) is used to compute the assignment probability at each iteration. The most likely assignment in each medium is considered. Let $r_1(q), r_2(q) \in R$ be the most likely assignments of peak q in media 1 and 2. The assignment $q \mapsto r$ is added into the master list of assignment if both following conditions are satisfied:

$$r_1(q) = r_2(q)$$

and

$$r_m(q) \neq r_m(k), \quad \forall k \in Q, \quad k \neq q.$$

It is possible that the second condition is not satisfied. In this case, a maximum bipartite matching is chosen.

Suppose that peak q has been assigned to r , then q and r are removed from the consideration of next iterations. At the end of each iteration, the alignment tensors S_1 and S_2 are refined based on the master list of assignments and model. At the same time, the assignment probabilities are renormalized and updated.

Time Complexity. It takes $O(n^2)$ time to construct the probability matrices. Updating the alignment tensors S_1 and S_2 takes $O(n^2)$. Since at least one residue is assigned at each iteration, there are at most $O(n)$ iterations. Thus, the total assignments are guaranteed to be finished in $O(n^3)$. The time complexity has been improved to be $O(n^{5/2} \log(cn))$ in [5], where c is the maximum edge-weight in an integer-weighted bipartite graph.

3 An Expectation/Maximization NVR Algorithm

An expectation/maximization NVR algorithm is applied in [1] for automated NMR resonance assignments. In the expectation/maximization framework, the observed variables include chemical shifts, d_{NNS} , amide exchange rates. The hidden variables include the correct resonance assignments and the correct alignment tensors. Let X be the set of observed variables, and Y be the set of hidden variables. Let Θ denote the set of all assignments obtained so far. Initially, Θ is set to be empty, and then grows as the algorithm finds more assignments. The Expectation (E) step of the EM algorithm computes the following expectation:

$$E(\Theta \cup \Theta' | \Theta) = E(\log \Pr(X, Y | \Theta \cup \Theta')),$$

where Θ' denotes the set of new assignments that is disjoint with previous set Θ . The Maximization (M) step computes the new assignment Θ^* that has the maximum likelihood:

$$\Theta^* = \operatorname{argmax}_{\Theta'} E(\Theta \cup \Theta' | \Theta).$$

At the same time, the algorithm updates the master list of assignment, i.e., $\Theta \leftarrow \Theta \cup \Theta^*$. Also, the alignment tensors are recomputed at the end of each iteration, based on updated assignments in Θ . The algorithm terminates when all peaks have been assigned.

4 3D Structural Homology Detection via NVR

Given unassigned RDCs of a target protein and known models in a database, we measure the 3D structure homology between the target protein and each known model by computing the likelihood of the assignment from the experimental RDCs to the model. We run the NVR algorithm for each model in the database, and then rank them by the likelihoods of the assignments.

Time Complexity. Suppose that there are in total k number of models in the database. Since the NVR algorithm runs in $O(n^{5/2} \log(cn))$ time, processing all k proteins takes $O(kn^{5/2} \log(cn))$. Finally, sorting all these k models according to their likelihoods of assignments takes $O(k \log k)$ time. Thus, the total required time is $O(kn^{5/2} \log(cn) + k \log k)$.

References

- [1] Christopher James Langmead and Bruce Randall Donald. An Expectation/Maximization Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *Journal of Biomolecular NMR*, 2004; 29(2):111-138.
- [2] Christopher James Langmead, Anthony K. Yan, Ryan H. Lilien, Lincong Wang, Bruce Randall Donald. Large a polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *RECOMB*, 2003: 176-187.

- [3] Christopher James Langmead and Bruce Randall Donald. 3D Structural Homology Detection via Unassigned Residual Dipolar Couplings. *Proc. IEEE Computational Systems Bioinformatics Conference (CSB)*, Stanford University, Palo Alto (August 10, 2003) pp. 209-217.
- [4] Christopher James Langmead and Bruce Randall Donald. High-Throughput 3D Structural Homology Detection via NMR Resonance Assignment. *Proc. IEEE Computational Systems Bioinformatics Conference (CSB)*, Stanford CA, (August, 2004) pp. 278-289.
- [5] Christopher James Langmead and Bruce Randall Donald. An Improved Nuclear Vector Replacement Algorithm for Nuclear Magnetic Resonance Assignment. *Tech. Rep. TR2004-494*, Dartmouth Dept. of Computer Science, 2004.
- [6] L. Wang and B. R. Donald. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *Jour. Biomolecular NMR*, 29(3):223-242, 2004.