Determination of macromolecular structure and dynamics from experimental data: NMR structure determination as an example

Michael Nilges
Unité de Bio-Informatique Structurale
Institut Pasteur
nilges@pasteur.fr
Overview

1. Introduction: relating data to structure
2. Hybrid energy and treatment of errors
3. Minimisation of hybrid energy
4. Relation to probability theory
5. Sampling probability densities
1. Relating data to structure and the hybrid energy concept
2. Hybrid energy and treatment of errors
3. Minimisation of a hybrid energy
4. Relation to probability theory
5. Sampling probability densities
Biomolecular structure determination

• Biomolecular structure determination includes a phase of molecular modeling,
  • fit a model to the data
• This is (one of) its most important applications
  • CNS (1998) citations 13806
  • CHARMm (1983) citations 8853
Why is fitting data difficult

- In particular for biological macromolecules, data are
  - incomplete
  - noisy
  - contradict each other
  - contradict prior knowledge
- Theoretical (or forward) models are
  - incomplete (parametric, with non-measurable parameters)
  - very approximative
Data are incomplete

- For biomolecules, number of parameters (coordinates) usually exceeds number of observables
- Number of degrees of freedom: 3N
- Number of observables
  - X-ray: number of reflections depends on resolution
  - NMR: number of NOEs etc: < 20/ aa
- Need to complement data with prior information
  - geometric: bond lengths, bond angles, planarity, vdW radii
  - force fields
Data are noisy

- The measurement of a quantity is not exactly reproducible
- Measurements follow a certain distribution
- Example:
  - Gaussian (normal) distribution of error for \( x \) around mean \( \mu \),
  - standard deviation \( \sigma \),
  - \( \Rightarrow \) probability is

\[
f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}
\]
Michael Nilges. Structure calculation from NMR data.
Data may contradict each other

- Example: NOEs of same side-chain to different positions
  - effect of dynamics
  - cannot be satisfied in one single structure
Data and forward models: the NOE

- Inter-proton distances can be derived from NOESY experiments
  - Ideally, we can measure the “cross-relaxation rate”
  - which depends on spectral densities
  - which depend on correlation functions (radial and angular fluctuations)
  - internal (local) dynamics can be separated from overall tumbling, simplify to two exponentials

$$\sigma_{ij} = \frac{\pi}{5} \gamma^4 \hbar^2 [6J_{ij}(2\omega) - J_{ij}(0)]$$

$$J_{ij}(\omega) = 2 \int_0^\infty C_{ij}(t) \cos(\omega t) dt$$

$$C_{ij}(t) = \frac{1}{4\pi} e^{-t/\tau_R} \left\langle \frac{P_2(\hat{\mu}_D(0) \cdot \hat{\mu}_D(t))}{r_{ij}^3(t) r_{ij}^3(0)} \right\rangle \frac{C_O(t)}{C_I(t)}$$

$$C_{I}^{LS}(t) = \left\langle r^{-6} \right\rangle (S^2 + (1 - S^2) e^{-t/\tau_e})$$
Data and forward models: the NOE

- Inter-proton distances can be derived from NOESY experiments
  - Ideally, we can measure the “cross-relaxation rate”
  - which depends on spectral densities
  - which depend on correlation functions (radial and angular fluctuations)
  - internal (local) dynamics can be separated from overall tumbling, simplify to two exponentials

\[
\sigma_{ij} = \frac{\pi}{5} \gamma^4 \hbar^2 [6J_{ij}(2\omega) - J_{ij}(0)]
\]

\[
J_{ij}(\omega) = 2 \int_0^\infty C_{ij}(t) \cos(\omega t) dt
\]

\[
C_{ij}(t) = \frac{1}{4\pi} e^{-t/\tau_R} \left\langle \frac{P_2(\hat{\mu}_D(0) \cdot \hat{\mu}_D(t))}{r_{ij}^3(t)r_{ij}^3(0)} \right\rangle_{C_O(t)} C_I(t)
\]

\[
C_{IS}^L(t) = \left\langle r^{-6} \right\rangle (S^2 + (1 - S^2)e^{-t/\tau_e})
\]
NOE and dynamics

\[ C_I^{LS}(t) = \left\langle r^{-6} \right\rangle \left( S^2 + (1 - S^2)e^{-t/\tau_e} \right) \]

\[ \sigma_{ij} \propto r_{ij}^{-6} f(\tau_c) \]

- If distance is rigid and known, NOE serves to measure local angular fluctuations
  - generalised order parameter \( S^2 \) characterises local fast dynamics
  - \( S^2 \) is 0 when completely flexible
  - \( S^2 \) is 1 when completely ordered
Forward model

- Obtain calculated crossrelaxation rate from structure / long dynamics trajectory
- Can we invert this?
  - how to impose crossrelaxation rate on trajectory?
  - how are crossrelaxation rate and NOE related?
    - “spin diffusion”
- Standard solution:
  - neglect dynamics (and spin diffusion) and treat as “noise”
  - try to obtain single structure from data
Example: distance measurement from NOEs, only protons

Forward model:
- isolated spin pair approximation
- NOE depends on distance (< 4 Å)

\[ \sigma_{ij} \propto r_{ij}^{-6} f(\tau_c) \]

\[ NOE_{ij} \propto r_{ij}(\mathbf{x})^{-6} \]

\[ r_{ij}^0 \approx (C_{cal} NOE_{ij})^{-\frac{1}{6}} \]

approximate model neglects
- internal dynamics
- spin diffusion

 calibration factor \( C_{cal} \) unknown (not measurable)

end result: approximate distances
Data may contradict prior information

- Consequence of forward model:
  - “one rigid structure” should satisfy data
  - other approximations in forward model
  - incorrect parameter choice
- Consequence of data:
  - noise
  - “false positives”
- Structure calculation is always a compromise between satisfying data and prior information
Hybrid Energy

- combine data and physical model of molecule force field into one function (target function, hybrid energy function)
  - X-ray: Jack & Levitt, 1978

$$E_{\text{hybrid}} = E_{\text{phys}} + w_{\text{data}} E_{\text{data}}$$

- guess $w_{\text{data}}$ : “something of a problem”
- minimise this function

- hybrid energy function complements incomplete data
- and prevents that structure deviates too much from expectation
1. Relating data to structure and the hybrid energy concept
2. Hybrid energy and treatment of errors
3. Minimisation of a hybrid energy
4. Relation to probability theory
5. Sampling probability densities
Some history

- combination with MD (for NMR refinement):
  - van Gunsteren, Kaptein 1985

- combination with MD (for folding):
  - Brunger, Clore 1986

- combination with MD (for X-ray refinement)
  - Brunger 1988
How to construct $E_{hybrid}$

- $E_{phys}$:
  - derive from standard force field
    - modifications may be necessary
    - more rigid
    - simplify non-bonded
  - derive from statistical analysis
    - derive from covalent parameters in small molecules
    - mean values
    - force constants from variation
NMR structure calculation: simplified force field

- covalent interactions: rigid, uniform force constants
- ideal values from Engh & Huber
- vdW interaction: quartic potential, soft core, no attractive part
- no electrostatics
How to construct $E_{\text{hybrid}}$

- $E_{\text{data}}$:
  - derive from distribution of data
  - e.g., for Gaussian distribution of distances from NOEs

$$P(D|X, \sigma) \propto \exp \left[ \frac{- (r - r(X))^2}{2\sigma^2} \right]$$

- the potential function would be

$$E_{\text{NOE}} \propto \sum_{i=1}^{N_{\text{NOE}}} \left( r_i(x) - r_i^0 \right)^2$$
Treat data and model imperfections in $E_{hybrid}$

- All data contain errors (experimental noise)
- All forward models contain approximations
- No ideal agreement between calculated and measured data possible
- Need to adapt $E_{hybrid}$
  - use appropriate weight
  - use appropriate functional form
Example: Flat-bottom-harmonic-wall (cheat)

\[ E_{\text{data}} \propto \sum_{i}^{N_{\text{noe}}} \begin{cases} 
(r_i(X) - L_i)^2 & \text{if } r(X) < L_i \\
0 & \text{if } L_i \leq r(X) \leq U_i \\
(r_i(X) - U_i)^2 & \text{if } r(X) > U_i 
\end{cases} \]

- loose upper and lower bounds
- FBWHH potential
  - flat bottom harmonic walls
  - no force between L and U
  - not optimal
Advantages:
- simple concept of geometrical consistency ("Distance Geometry")
- weight is not important
- fast

Disadvantages
- not derived from error distribution
- human bias: where to put the upper bound?
- false sense of "security"
- loss of information

FBHW potentials are a bad good idea
Adapt weight

- **Empirical methods**
  - “experience”
  - adjust average gradients (Jack & Levitt):
    - Ephys and Edata have equal importance
  - cross-validation (Brunger)
    - divide data into “test set” / “working set”
    - use only working set for calculation
    - use only test set for evaluation
    - look for minimum or elbow region

- **Bayesian methods**
Overview

1. Introduction: relating data to structure
2. Hybrid energy and treatment of errors
3. Minimisation of hybrid energy
4. Relation to probability theory
5. Sampling probability densities
Minimisation algorithms for NMR

- Energy minimisation
- Simulated annealing
  - Molecular dynamics
  - Torsion angle dynamics
  - (Monte Carlo)
- Distance geometry
- Genetic algorithm
- ...

Michael Nilges. Structure calculation from NMR data.
Multiple minimum problem

High energy barriers to fold protein

Standard minimisation only "downhill"
Minimisation by molecular dynamics

\[ \frac{d^2 r_i}{dt^2} = - \frac{c}{m_i} \frac{\partial}{\partial r_i} E_{\text{hybrid}} \]

- Molecular dynamics solves Newton's equations of motion
- Molecular dynamics can overcome local energy barriers
Newton dynamics

- MD is a minimiser with memory:
  - direction of motion depends on
  - force (derived from force field and experimental restraints)
  - momentum
Temperature control and variation: "MD-simulated annealing"
Energy scaling

- more flexible annealing schemes
- different variation of different energy terms
- equivalence:
  - mass/ energy/ temperature scaling

\[ \frac{d^2 r_i}{dt^2} = - \frac{c}{m_i} \frac{\partial}{\partial r_i} E_{\text{hybrid}} \]
Michael Nilges. Structure calculation from NMR data.

![Graph showing the relationship between temperature, physical energy ($E_{phys}$), covalent energy ($E_{covalent}$), van der Waals energy ($W_{vdW}$), and nuclear Overhauser effect ($W_{NOE}$) over time (steps). The graph illustrates the cooling process from search to cool1 to cool2 at 2000 K temperature.]
Michael Nilges. Structure calculation from NMR data.

- **start**
- **high temperature**
- **cooling and minimization**
Structure calculation with MD

- NMR data: distances
- Start: random structure
- Difficult search problem: many degrees of freedom
Structure calculation with MD

100 atoms:

1988:

20000 s per structure on mainframe (DISGEO, Havel)

now:

20 s per structure on PC
Torsion angle dynamics

- dynamics time step dictated by bond stretching: waste of CPU
- important motions are around torsions
- ~ 3 degrees of freedom per AA (cf 3Natom for Newton dynamics)
- Available in X-PLOR, CYANA, CNS, X-PLOR-NIH, ISD

\[ M(\phi) \frac{d^2 \phi}{dt^2} + C\left( \frac{d\phi}{dt}, \phi \right) = 0 \]
Calculation of structure ensembles

- with identical data/restraints:
  - repeat calculation (20-100 times)
  - random variation of initial conditions (starting structure/velocities)
- poor man’s “probability distribution”
  - obtain information on uniqueness / different fold
Meaning (?) of structure ensembles

• Simple way to assess uniqueness of solution
• This has very little to do with dynamics
• Distribution depends on
  • data
  • data representation
  • algorithm
  • forcefield
  • algorithm parameters
  • ...
So what about dynamics

- all data represent ensemble (and time) averages
- qualitatively, ensembles can show features of real dynamics
- why?
  - NOE distance potential resembles elastic network
  - NOEs can be absent because of dynamic effects
So what about dynamics

- depends very much on fold

PH domain

dsdb protein
NMR: rich source dynamic information

- experimental data sensitive to structure and dynamics (time and ensemble averages)
- example NOE:
  - inconsistencies in the derived distances
- spin relaxation experiments: fast ps dynamics
  - dipole–dipole interactions between 15N and H
- slower dynamics:
  - RDCs, J-couplings, chemical shifts, relaxation-dispersion
- dynamics measurement do not give atomic picture of motion
  - need models (e.g., MD simulations)
methods to include dynamics in calculation

- ensemble averaged restraints:
why is this so difficult

- bad observable to parameter ratio:
  - NOE data not sufficient to determine structure for ensemble
    - statistical relevance (only Bonvin & Brunger, with cross validation)
    - is the ensemble / trajectory Boltzmann distributed ?
  - how to keep ensemble together ? force field / ad hoc potential
    - contributions from data ? methods rely heavily rely heavily on quality of MD force fields, other experimental information; ad hoc potentials
  - quantitative data treatment (in particular for NOEs)
    - structures not determined from NOE data
    - exception: Vögeli et al., with “exact” NOEs
- mixture of different time scales
  - “order parameter” : ps time scale
  - RDCs, NOEs: ns - μs time scale
standard NMR ensemble and EROS ensemble

1. Introduction: relating data to structure
2. Hybrid energy and treatment of errors
3. Minimisation of hybrid energy
4. Relation to probability theory
5. Sampling probability densities
Minimisation and probability

- Where do potential forms come from
- Where do all the parameters come from
  - bounds
  - weights
  - any parameter required by theory
Probability and energy

\[ E_{\text{hybrid}} = E_{\text{phys}}(X) + \omega_{\text{data}} E_{\text{data}}(D, X) \]

- force field \( E_{\text{phys}} \) \( \Leftrightarrow \) probability (Boltzmann)
- probability of distortion of molecule
- force field: background information \( I \)
- prior probability

\[ P(X|I) = \exp \left[ - \frac{E_{\text{phys}}(X)}{kT} \right] \]
Probability and energy

\[ E_{\text{hybrid}} = E_{\text{phys}}(X) + w_{\text{data}} E_{\text{data}}(D, X) \]

- similar: \( E_{\text{data}} \leftrightarrow \text{probability} \)
- probability that data is correct, given structure \( X \):
- “likelihood”
Likelihood and restraint potential

- Inversely, if we know probability distribution, we can derive potential

\[ E_{data} \propto -\log [P(D|X, \sigma)] \]

- For Gaussian error, harmonic potential ("least squares")

\[ E_{data} \propto \frac{1}{2\sigma^2} (r - r(X))^2 \]

- The weight is related to the error in the data
Distances (NOEs) do not follow Gaussian

Gaussian distribution of logarithms

Gaussian distribution

Rieping, Habeck, Nilges, JACS 2005
Log-normal distribution

- Log-normal distributions
- and derived potentials

\[
\text{LN}(x_0, x, \sigma) \equiv \frac{1}{\sqrt{2\pi\sigma^2x_0}} \exp\left[-\frac{1}{2\sigma^2} (\log[x_0] - \log[x])^2\right]
\]
Several interesting properties:
- Only one free parameter (weight)
- Shape does not change with exponential ($d, d^{-6}$)
- “Flattening” of potential for inconsistent distances
Joint probability from prior and likelihood

- To calculate joint probability from single probabilities, multiply:

$$P(X|D, I) \propto P(X|I)P(D|X, \sigma, I)$$

Probability of a structure:
Posterior Probability

Bayes

Laplace
Hybrid energy revisited

\[ E_{\text{hybrid}} = E_{\text{phys}}(X) + w_{\text{data}} E_{\text{data}}(D, X) \]

- the hybrid energy function is negative logarithm of joint probability
- minimum energy corresponds to maximum probability
- data weight “should” depend on data quality
- story is incomplete (what about \( w_{\text{data}} \)?)

Probability of a structure

\[ P(X|D, I) \propto P(X|I)P(D|X, \sigma, I) \]

prior distribution
Hybrid energy and Bayesian probability

\[ E_{hybrid} = E_{phys} + \omega_{data} E_{data} \]

\[ p(X, \sigma | D, I) \propto \pi(X | I) \pi(\sigma | I) L(D | X, \sigma, I) \]
Bayesian determination of data weight

• Bayesian analysis:
  • Extended hybrid energy function including data quality
    \[ E_{\text{hybrid}}(X) = E_{\text{phys}}(X) + w_{\text{data}} E_{\text{data}}(X) + E_{\sigma}(X) \]
  • weight \(\leftrightarrow\) overall data quality
  • average weight over all structures
• Minimisation does not result in vanishing weight
• Update weight iteratively during structure calculation
• Meaningful estimate of data quality from chi2

comparison to X-ray for LogNormal

BPTI, 0.62 Å

IL8, 1.44 Å

GB1, 0.5 Å

IL4, 1.14 Å
PCA analysis for IL4

![PCA Analysis Chart]

- FBHW
- LogNormal
- X-ray

**Axes:**
- EV1
- EV2
1. Introduction: relating data to structure
2. Hybrid energy and treatment of errors
3. Minimisation of hybrid energy
4. Relation to probability theory
5. Sampling probability densities
Problems inherent in minimisation

- data incomplete: solution is degenerate
- data inconsistent: no solution exists
- many unknown parameters ("nuisance parameters")
- sparse data:
  - problems with determining auxiliary parameters
  - structure calculation difficult
- no objective figures of merit for structures
  - RMSDs and R-factors depend on all auxiliary parameters
  - few restraints can change result drastically
  - no concept to evaluate data quality ("don’t overfit"...“use data not used in structure calculation”...)
More rigorous modelling: probabilistic view

(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...

HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS

http://www.zeably.com/Bayesian_statistics
Probabilistic view of structure determination

- Evaluate probability for “all” conformations \( X \)
  - and all other parameters necessary for description of problem
- Prior information: energy of a conformation
  - use “weak prior”: covalent geometry / elastic network and “soft spheres”
- Likelihood / “satisfaction of restraints”
  - difference between exp. data and predicted data (forward model)
  - contains unmeasurable quantities (e.g., data quality \( \sigma \), theory parameters \( \xi \))
- Bayes’ theorem:

\[
P(X, \sigma, \xi | D, I) \propto P(X | I)P(\sigma)P(\xi)P(D | X, \sigma, \xi)
\]
Probabilistic view of structure determination

- Evaluate probability for “all” conformations $X$
  - and all other parameters necessary for description of problem
- prior information: energy of a conformation
  - use “weak prior”: covalent geometry / elastic network and “soft spheres”
- likelihood / “satisfaction of restraints”
  - difference between exp. data and predicted data (forward model)
  - contains unmeasurable quantities (e.g., data quality $\sigma$, theory parameters $\xi$)
- Bayes’ theorem:

\[
P(X, \sigma, \xi | D, I) \propto P(X | I) P(\sigma) P(\xi) P(D | X, \sigma, \xi)
\]
Bayesian structure determination

• “Inferential Structure Determination” (ISD)
  • Michael Habeck, Wolfgang Rieping (Rieping et al., Science 2005)
  • re-determine forward models for each data type
  • calculate densities (not single minimum structures) for
    • all unknowns (including but not exclusively the coordinates)
  • their uncertainty with interdependencies

\[
P(X, \sigma, \xi | D, I) \propto P(X | I) P(\sigma) P(\xi) P(D | X, \sigma, \xi)
\]
Forward model and likelihood in NMR

- Forward model to evaluate likelihood
  - simple functional form: \( \text{NOE} \propto r^{-6} \)
  - includes all unknown parameters (e.g., data quality \( \sigma \))
  - data weight depends on \( \sigma \) and is unknown
  - NOEs and derived distances: log-normal distribution
    - Rieping, Habeck, Nilges, JACS 2005

\[
\text{LN}(x_0, x, \sigma) \equiv \frac{1}{\sqrt{2\pi \sigma^2}} \frac{1}{x_0} \exp\left[ -\frac{1}{2\sigma^2} (\log[x_0] - \log[x])^2 \right]
\]
Sampling probability distributions

- Posterior $P(X, \sigma, \xi | D, I)$ is very complex
  - too many degrees of freedom for systematic search
  - coordinates $X$ and other parameters $\sigma, \xi$
- Representative samples
- Frequency $\propto$ probability
- Markov Chain Monte Carlo
  - detailed balance

\[
P(X, \sigma, \xi | D, I) \propto P(X | I) P(\sigma) P(\xi) P(D | X, \sigma, \xi)
\]
Sampling probability distributions

• Computationally complex
  • cf calculating partition function in statistical mechanics

• ISD algorithm uses
  • hybrid Monte Carlo (HMC)
  • replica exchange
  • Tsallis distribution
  • Gibbs sampling for additional parameters
  • Habeck, Rieping, Nilges, Phys Rev Lett 2005

• ISD outperforms standard structure calculation in NMR

• HMC-replica exchange is inefficient, problem for large systems
• Test other sampling algorithms (NCMC)
Sampling nuisance parameters

Gibbs sampling of nuisance parameters ($\{\theta\}$ fixed):
$\gamma, \sigma \sim p(\gamma, \sigma | \{\theta\}, D, I_0)$

Hybrid Monte-Carlo sampling of internal coordinates ($\gamma, \sigma$ fixed):
$\{\theta\} \sim p(\{\theta\} | \gamma, \sigma, D, I_0)$

$q_1, q_2, \ldots, q_n$

Replica MC: Simulation of $n$ independent Tsallis-transformed posterior distributions

- data quality $\Leftrightarrow$ weight
- scale factor
- other parameters

(not assumed known)

(usually determined by empirical methods: experience, crossvalidation)
Typical trace (SH3 domain)

- replica exchanges
- “energy”
- data variance
- calibration
Distribution of $\sigma$ in Ubiquitin and SH3

- Distributions for all parameters
- No fixed "weight" but distribution
  - "marginalization": integration over all other parameters
    - coordinates
    - scale factor

\[
P(\sigma \mid D,I) = \int d\theta d\gamma P(\sigma \mid \theta, \gamma)P(\theta,\gamma \mid D,I)
\]
Computational requirements

- a few days on 50 Linux PCs
  - every “supertransition” is 50 short dynamics trajectories
  - in total, > 25000000 hybrid Monte Carlo steps
  - convergence of distribution, not only structures
in the practical we will use NCMC trajectories

- **Sampling:** nonequilibrium candidate MC (NCMC)
  - alternate small and large random perturbations of phi and psi
  - relaxation (200-1000 steps NVE molecular dynamics)
    - simplest version of method by Nilmeier et al. & Chodera (2012) PNAS
    - implementation in CNS
• perturbation:
  • random rotation about a fraction of the torsion angles
  • choose optimal weight for the obtained coordinates (no sampling of w_data)
  • add compensation term to the total energy

• relaxation:
  • NVE MD trajectory (250-750 steps)

• Metropolis criterion
  • total energy differences (Ephys + Edata + Ekin + Esigma)
  • between trajectory endpoints

• replica exchange (two replicae)
• short trajectories
Summary

- Minimising hybrid energy corresponds to maximizing the probability of a structure, given data and force field
- ...if one knows the data quality, scale factors, ...
- Relationship of error distribution and restraint potentials
- Weights on data terms
  - usually set empirically (trial and error, experience, cross validation)
  - Bayesian determination of weight possible
- Bayesian probability theory:
  - theoretical foundation of structure refinement
Summary (Bayes)

- Joint probability distributions can be determined by sampling
  - sampling: frequency is proportional to probability
  - e.g., Markov-Chain Monte Carlo methods
  - hybrid Monte Carlo has advantages for coordinates
- All unknown parameters can be sampled
  - coordinates
  - parameters of the forward model
  - quality of data <= weights on data
- Distributions of parameters of interest can be obtained by marginalization


• Güntert P. Automated NMR structure calculation with CYANA. Methods Mol Biol. 2004;278:353-78


• Markwick PR, Nilges M.
Literature: Bayesian

- Rieping W, Habeck, M, Nilges, M (2006). Refinement against NOE intensities using a lognormal distribution improves the quality of NMR structures. JACS,