

LECTURE NOTES FOR INTERFERENTIAL STRUCTURE DETERMINATION

(Intended as a loose introduction)

Consult the original literature

1) The ISD manual from www.isd.bio.cam.ac.uk

2) References therein, in particular:

a. W. Rieping et al. Science 309, 303 (2005)

b. M. Habeck et al. PRE 72, 031912 (2005)

c. M. Habeck et al. PNAS 103, 1756 (2006)

- DETERMINATION OF MACROMOLECULAR STRUCTURES FROM EXPERIMENTAL DATA IS AN ILL-POSED PROBLEM
(Less data than unknown)
- IN THE STANDARD APPROACH (see discussion of 'Integrative structural biology' in the previous lecture) IT TAKES THE FORM OF AN INVERSE PROBLEM
- REMARK: Inverse problems are not new for atomistic simulators; see - e.g. Henderson's theorem (1974) stating - in a nut-shell - that: If two pairwise additive potential functions produce the same radial distribution function then they differ by a constant
(ask Giuseppe D'Adamo for illustration)
- In the ISD structure determination is seen as an inferential problem (i.e. take a decision under uncertainty) based on posterior probability evaluation using Bayes' theorem, leading to a fair estimate of uncertainties (e.g. 'error bars' on the obtained structure)
- REMARK 2: The very concept of a structure represented by $X \in \mathbb{R}^{3N}$ is far from trivial (expand)

ISD continued

- STRUCTURAL DETERMINATIONS ARE COMMONLY ASSOCIATED TO X-ray CRISTALLOGRAPHY OR NUCLEAR OVERHAUSER EFFECT (NOE, Intensity change of the intensity of a NMR resonance due to supersaturation (pumping) of a nerby resonance due to a close nuclear spin $I \sim \gamma/r^6$)
- THE DATA ARE DIFFRACTOGRAMS OR NOE maps (2D data, spots on a plane)
- LET US DENOTE MEASUREMENTS AS $\{y_i\}$ A DISCRETE SET OF EXPERIMENTAL NUMBERS; ANALYSIS OF DATA REQUIRES A MODEL f , A 'FORWARD' MODEL THAT CONNECTS A MOLECULAR STRUCTURE x , PERTURBED BY THE MEASUREMENT, WITH A RESPONSE y

$$y = f(x)$$

- AS SAID ABOVE, STRUCTURE DETERMINATION IS AN ILL-POSED PROBLEM (EMPIRICALLY, THE SAME DATA CAN BE EXPLAINED BY MULTIPLE CONFORMATIONS, STRUCTURES)
- REM: UNDER THESE CONDITIONS, DECIDING ABOUT A UNIQUE STRUCTURE IS AN OVER FITTING PROCEDURE

- MOREOVER, EXPERIMENTAL DATA ARE
 - INCOMPLETE (LESS THAN NEEDED, SPARSITY)
 - NOISY (OF LIMITED REPRODUCIBILITY)
 - COME FROM DIFFERENT SETTINGS (HETEROGEN.)
- THE MODEL f ITSELF IS IN GENERAL APPROX.
(e.g. IN THE CASE OF NOES THE $I \approx \gamma f^2$ is based on the assumption of INDEPENDENT SPIN POLARIZATION)
- NEVERTHELESS, IN THE STANDARD (NON BAYES) OPTIMIZATION APPROACH ONE MAKE ATTEMPTS AT NUMERICALLY SOLVING THE INVERSE PROBLEM TO GET A UNIQUE SOLUTION X^*

(formally) $X_i^* = f^{-1}(y_i)$

- IN THE STANDARD APPROACH ONE INTRODUCES A HYBRID ENERGY FUNCTION

$$E(X, a) = E_\theta(X, a) + w_{\text{data}} E_{\text{data}}(X, a)$$

TO BE WORKED OUT AIMING AT AN OPTIMAL, MINIMIZING SOLUTION X^*

REM: $E_\theta(X, a)$ is a physical model, like a force field, a potential mechanical energy depending on a set θ of parameters and on the information about composition, chemistry of the molecule a . w_{data} is an adjustable scaling parameter; $E_{\text{data}}(X, a)$ represent the deviation, the degree of matching between data and X , given the model.

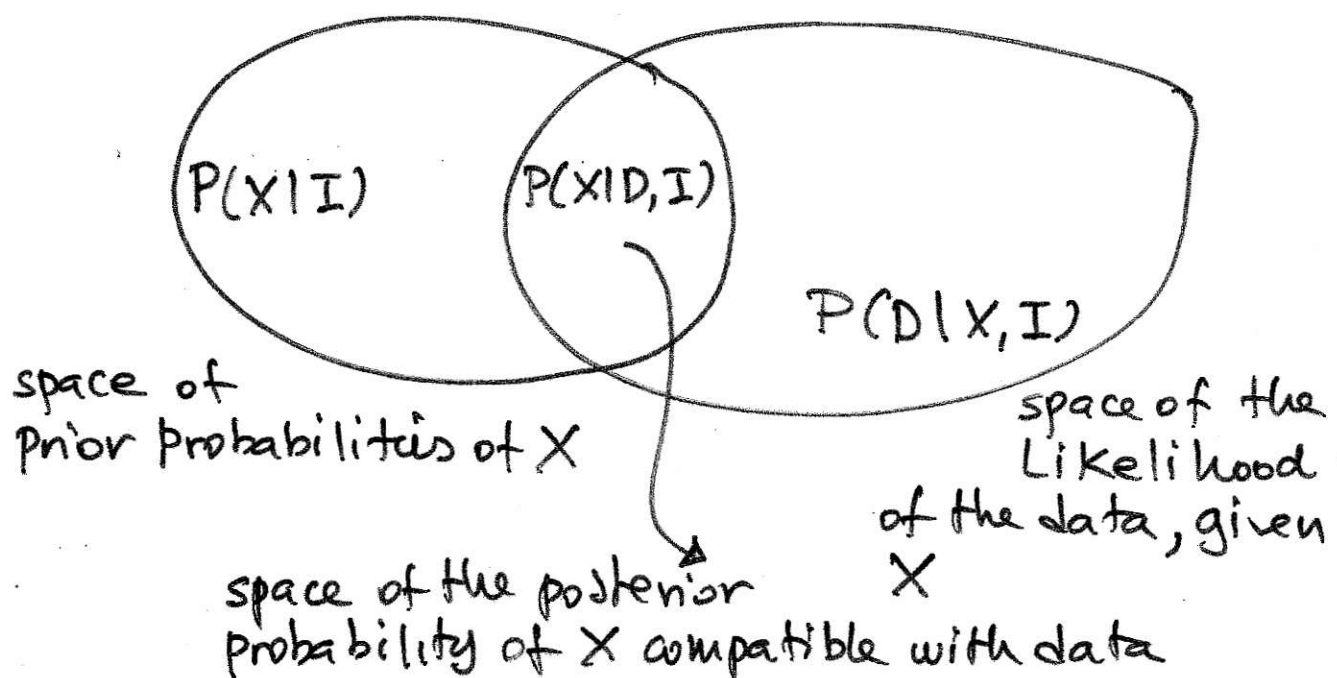
- IN THE STANDARD APPROACH $E_{data}(X, a)$ measures the fulfillment of constraints (review step 2 in integrative structure determination discussed in the previous lecture)
- In ISD structure determination is related to a ranking of structures X_i through numbers P_i , such that:

if $P_i > P_j$ then structure X_i is more compatible than X_j with experimental data. P_i 's, in a Bayesian setting, are so called POSTERIOR PROBABILITIES

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

(POSTERIOR) (LIKELIHOOD) (PRIOR)

- The 'essence' of Bayes' theorem is embodied in this Venn's diagram



In the ISD algorithm (see www.isd.bio.cam.ac.uk) arguments have been given to assume that the Likelihood has the form of a product of Lognormal distributions (pointing to multiplicative noise):

$$P(D|X, \alpha, \sigma, I) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2\sigma^2} \left[\log I_i - \log(\alpha d_i^{-6}) \right]^2 \right]$$

α, σ are the so called "nuisance" auxiliary parameters, not accessible to direct exp. determin.

- Then we have to figure out what $P(X|I)$ would be; reasonably one could get the Canonical ensemble:

$P(X|I) \propto \exp[-\beta E(x)]$, where $E(x)$ is a molecular force-field (in the simplest case, think about springs and balls model.)

- Assuming that α and σ are uniformly distributed

$$P(X, \alpha, \sigma | D, I) \propto \alpha^{-1} \sigma^{-(n+1)} \exp \left[-\beta E(x) - \frac{1}{2\sigma^2} \sum_i \left[\log I_i - \log(\alpha d_i^{-6}) \right]^2 \right]$$

And this is the distribution to be sampled

REM Q. What does I means in the notation $P(X|D, I)$?

A. I means information, The state of information, the assumptions made on prior knowledge to connect data with structures, in the present case

- In general, following the subjective approach of DeFinetti and Savage any probability one speaks about is a conditional probability, conditioned by a state of information

• RESUME

To take a decision about a structure X as inferred from data D we have to explore the distribution of $P(X|D, I)$ as the product LIKELIHOOD \times PRIORS

So we have to discuss a little, and in general how to get $P(D|X, I)$ and $P(X|I)$

- REM: The knowledge of a given structure X does not determine the intensity of a NOE peak \propto the association of $\gamma d^{-6} \rightarrow I_i$ is based on the assumption of independent spins; the spots' intensity is also affected by diffusive motions that blur the mapping. And, there are finite size effects, signal to noise problems i.e. experimental uncertainties and fluctuations

- NOW, THE POSTERIOR PROBABILITY $P(X, \alpha, \sigma | D, I)$ IS NUMERICALLY SAMPLED, THAT IS A SET OF X_s IS EXTRACTED FROM THE DISTRIBUTION

$X^{(k)}, \alpha^{(k)}, \sigma^{(k)}$ k -th iterated samples

from a Gibbs sampling scheme and a hybrid Monte Carlo Method (see literature)

Based on a replica-exchange MC method and on the use of Tsallis's statistics to modulate $P(X|I) \approx P(X, q | I)$

$$P(X, q | I) = [1 + \beta(q-1)] [E(x) - E_{\min}]^{-\frac{q}{(q-1)}}$$

for $q=1$ The canonical ensemble is recovered
NOTE

for $q > 1$ the suppression of high energy conformations is not exponential but close to power law so it is an effective way of doing a simulated annealing embodied in the variation of q , instead of a schedule of heating-cooling cycles.

EXPAND on TSALLIS STATISTICS