Modeling biomolecules in solution
pitfalls and challenges

Sylvia McLain

25th April 2013
Accuracy in Powder Diffraction IV
What is biology?

**Biology**
The study of living organisms, divided into many specialized fields that cover their morphology, physiology, anatomy, behavior, origin, and distribution
the plants and animals of a particular area as in “the biology of the Chesapeake Bay”
the physiology, behavior, and other qualities of a particular organism or class of organisms: “human biology”

**Molecular biology**
The branch of biology that deals with the structure and function of the macromolecules (e.g. proteins and nucleic acids) essential to life.

**Structural biology**
branch of molecular biology which is concerned with macromolecular structure and how this effects function
Challenge 1: Convincing biologists understanding structures really is biology

First published structure of globular myoglobin 1958
- Kendrew

First structure of myoglobin at 5.5 Å resolution
- Perutz

Insulin 1968 at 2.8 Å resolution
- Hodgkin
Protein crystallography data base (PDB)
- In 1972 there were 2 structures in the PDB
- 89,740 structures in the PDB (many are repeats)
- 18 are from Powder diffraction (X-ray)
- number of known *human* proteins estimated at 50,000
- most biological molecules do not crystallize
Challenge 2: Making protein powder diffraction a viable tool for structural biologists

Challenge 3: Crystallography (no matter how accurate) doesn’t always work

Most biological molecules don’t crystallize

Many biological molecules don’t crystallize as single molecules - complexes

Many biological molecules crystallize with a high level of disorder (disordered loops)

Most biological molecules are disordered *in vivo*

Even when biological molecules crystallize, they may not be the same in solution
Measuring biological molecules in solution

Role of water in biology not well understood

- role in ligand binding?
- role in association (membranes, protein folding, amyloid fibers)
- hydrophobic/hydrophilic forces
- 'oil and water don’t mix’ but water crosses membranes!
Measuring biological molecules in solution

In real life water is always around

- Protein folding
- Protein/peptide association
- Membrane formation
- DNA transcription
- Receptor ligand binding interactions
Challenge 4

Challenge 4: no Bragg scattering, disordered systems

protein crystallography

protein powder crystallography

Liquid diffraction
Neutron diffraction with isotopic substitution

- $F(Q)$ - structure factor
- $F(Q) = \sum c_\alpha c_\beta b_\alpha b_\beta S_{\alpha\beta}(Q)$

$b$ - neutron scattering length

- Different isotopes scatter with different intensity
- Measurement of chemically equivalent isotopically unique samples
- model with EPSR

S Busch,* et al., manuscript submitted (2013)
Empirical Potential Structural Refinement - computer modeling

\[ U_{\text{intra}} + U_{\text{inter}} \rightarrow \text{initial configuration} \]

\[ \downarrow \]

Monte Carlo Simulation

\[ \downarrow \]

\[ S(Q) - 1 = \frac{\sum_{\alpha \beta} c_{\alpha} c_{\beta} b_{\alpha} b_{\beta} (S_{\alpha \beta}(Q) - 1)}{\sum_{\alpha} (c_{\alpha} b_{\alpha})^2} \]

\[ \downarrow \]

\[ \Delta S_{\alpha \beta}(Q) = S_{\alpha \beta}^{\text{exp}}(Q) - S_{\alpha \beta}^{\text{sim}}(Q) \]

\[ \downarrow \]

\[ U_{\text{emp}}^{n+1} = U_{\text{emp}}^{n} + k_B T \cdot F[\Delta S(Q)] \rightarrow \]

\[ \downarrow \]

\[ g(r) \]

- Model specifically designed for amorphous systems
- Fits a set of neutron data
- Structural model only!
Dipeptides as a model system

Series of soluble peptides with increasing hydrophobicity
Very soluble in pure water
Can use H/D substitution on H atoms
Measured at high concentrations (2.5 M)
Association between peptides in solution

Peptide-peptide correlations

- $g_{OcHx}(r)$ coordination highest in gly-al a (0.74)
- Hydrophobic coordination highest in gly-al a (least hydrophobic peptide)

Association of peptides in solution

Clustering analysis of peptides in solution

Charge-charge interactions drive association in solution

Open symbols MD, closed symbols EPSR
Challenge 4: Most functional biological molecules are larger than dipeptides

- higher level of complexity
- lower level of solubility (less material in solution)
- disorder, disorder, disorder

Even slightly larger molecules start to cause problems
glycyl-L-prolyl-glycinamide in water

- Protein folding model
- \(\beta\)-hairpin turn motif
- role of water in folding initiation?

S Busch,* et al., manuscript submitted (2013)
Liquid diffraction - what it really looks like

What you really measure

\[
\frac{d\sigma}{d\Omega} = \frac{d\sigma}{d\Omega_s} + \frac{d\sigma}{d\Omega_d} = \sum \alpha b_\alpha^2 + P(Q, \theta) + F(Q)
\]
gpg - EPSR fit to the data

- box of molecules at $\rho$, $T$, $P$ of measurement
- reasonable R factor ($R_f$)

$$R_f = \frac{1}{M} \sum_i \left\{ \frac{\sum_Q (D_i(Q) - fit_i(Q))^2}{N_Q(i)} \right\}^*$$

- $M$ - number of data sets
- $D_i(Q)$ - data at point $Q$,
- $i$th data set
- $N_Q(i)$ - number of points in $Q$

*deliberately ignores statistical errors
systematic effects unknown
Peptide bonds not flat! Need additional constraints Need additional measurements
two average conformations in solution
cis vs trans with respect to the glycyl-prolyl bond
gpg - which fits are better?

before molecular constraint

after molecular constraint
gpg - which fits are better?

before molecular constraint

after molecular constraint
- unphysical distances
- $\text{Na}^+$ association?
- hard to detect
- EPSR perfectly 'happy'
gpg - which fits are better?

before reduced charges

after reduced charges
gpg - which fits are better?

before reduced charges

after reduced charges
Challenge 5: Building consistent models

Liquid diffraction data on its own for complex systems not enough

- more neutron diffraction data sets (more isotopic substitutions)
- add X-ray diffraction
- NMR
- other simulation techniques - MD, DFT (other?)
water-mediated turning in gpg - the consistent result

consistent evidence

diffraction
NMR - chemical shifts
EPSR simulations
MD simulations
Challenge 6

Modeling constrained by more experimental data

- improving reverse Monte Carlo methods
- more data to be 'fit' such as NMR
- using potentials from EPSR to inform MD
- link structure with dynamics

Challenge 6b: Moving towards more complex systems

- boundary between polycrystalline systems and amorphous systems
- 'disordered refinement' - PDF for biomolecules
- Spanning from the Å scale to macromolecular scale
- New neutron instrumentation - NIMROD and SANS2d at ISIS (UK)
Acknowledgments

Oxford

Sebastian Busch
Andrew Johnston
Christina Redfeild

Alan Soper (ISIS)
Chris Lorenz (King’s College London)

Funding

UK Engineering and Physical Sciences Research Council
Wellcome Trust