Predicting Structure and Function: from dreams to reality

Janet Thornton; EBI
Predicting Structure and Function: from my dreams to my reality

Janet Thornton; EBI
The unique 3D structure of a protein determines its function.

During evolution, structure is much better conserved than sequence.
Papers with Protein Structure Prediction in Title

Number of papers

Decade

70s  80s  90s  00s  2010-2012
Ab initio Prediction, using semi-empirical energy functions and energy minimisations.

Minimisation of AppA, an analogue of NAD

Conformational Energy Calculations for Dinucleotide Molecules. A Systematic Study of Dinucleotide Conformation, With Application to Diadenosine Pyrophosphate

JANET M. THORNTON* and PETER M. BAYLEY, Biophysics Division, National Institute for Medical Research, Mill Hill, London NW7 1AA, U.K.

Synopsis

A systematic study of the conformational states of the dinucleotide diadenosine 5',5'-pyrophosphate (AppA), an analog of the coenzyme NAD⁺, has been made using semi-empirical energy calculations. Taking low-energy mononucleotide structures as starting conformations, energy minimizations have been performed. The most stable structures exhibit stacking interactions between the adenine bases; there are many different stacked states of similar energy; their stability is derived from nonbonded interactions primarily between the bases but also from base-sugar interactions. The most common form of stacking in the most stable structures was found to be antiparallel A-A helix. These findings are consistent with the experimental data, which suggest that AppA adopts predominantly a stacked state in solution, and this state incorporates a variety of stacked conformations.

The conformational properties of purine and pyrimidine mononucleotides have been studied by a variety of theoretical and experimental techniques. In a previous paper, we presented a relatively simple semi-empirical approach to the theoretical determination of the allowed conformations of 5'-AMP, NMN⁺, and 3'-AMP. This approach is readily extended for a treatment of dinucleotides. The present paper describes calculations on diadenosine pyrophosphate, AppA, a 5',5' di-nucleotide related to the nicotinamide adenine dinucleotide coenzyme NAD⁺.

Experimental work, especially optical and nmr spectroscopy, has revealed that in solution the dinucleotides may exist in folded conformations with stacking interactions between the bases. Thus both NAD⁺ and AppA exhibit hypochromism, positive "couplet" circular dichroism, ring current shifts, and temperature, pH, and solvent dependence.

* Present address (and to whom proofs should be sent): Laboratory of Molecular Biophysics, Department of Zoology, Oxford OX1 3PS, U.K.

** Abbreviations: AppA: diadenosine 5',5'-pyrophosphate, ApA: adenylyl 3',5'-adenosine, NMN⁺: nicotinamide mononucleotide.
Ab initio prediction for NAD – Lessons learnt

- Energy Functions far from perfect
  - VDW radius + 0.2 angstroms needed to predict correct base stacking
  - Electrostatics approximate!

- 10 variables too much to minimise by hand
  - (c.f. phi/psi for 5 amino acids)

- Energy minimisation an art!
  - Steepest descent/conjugate gradient methods rapid, but gets stuck in local minimum
  - Nelder & Mead (Simplex approach) – explores larger space but slow
  - Solution is to alternate

- Folding Pathway important to avoid local minima (in silico)
  - Start in the middle for NAD

- Energy minimum is not necessarily biologically relevant!
NAD binds to Lactate Dehydrogenase in extended conformation.
PDB in 1970s

Year

Number of structures
1 6 9 14 58 85 104 128 150

Total Entries


Year

Number of structures
1970’s
Basic Principles of Protein Structure

- Properties of amino acids eg helix propensities, Allowed phi, psi values
- Hydrophobic Core
- Secondary Structures Geometries
- Chirality; sheets, βαβ, Barrels
- Structure Annotation & visualisation
Prediction

- Secondary Structure prediction
  - Many different methods
    - C&F; Information Theory; Lim
  - Accuracy crept up slowly (60-70%)

- Dynamics of folded proteins:
  - JA McCammon,
    BR Gelin & M Karplus
    Nature 267, 585 - 590 (1977)

- Modelling by hand

Cheating!
review article 1978

Prediction of protein structure from amino acid sequence

Michael J. E. Sternberg* & Janet M. Thornton*

Nature 271, (1978)

Protein Folding Data might help!
1980’s
Protein Tertiary Structures & Evolution;

- Interactions:
  - Amino acid packing
  - Tertiary packing – helix; sheet
- Domains
- Folds & architectures
- Evolution – conserved structures
Knowledge-based prediction of protein structures and the design of novel molecules

T. L. Blundell, B. L. Sibanda, M. J. E. Sternberg & J. M. Thornton

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Prediction of the tertiary structures of proteins may be carried out using a knowledge-based approach. This depends on identification of analogies in secondary structures, motifs, domains or ligand interactions between a protein to be modelled and those of known three-dimensional structures. Such techniques are of value in prediction of receptor structures to aid the design of drugs, herbicides or pesticides, antigens in vaccine design, and novel molecules in protein engineering.
Prediction in 1980s

- Many New Tools
  - Homology Modelling; Greer (1981) JMB 153
  - Simulations
  - Electrostatics
  - BLAST

Many more sequences: Dayhoff
Protein superfamilies and domain superfolds

Christine A. Orengo, David T. Jones & Janet M. Thornton
Prediction in 1990s and beyond

- **Secondary Structure prediction**
  - Major improvement with PHD and inclusion of homologous sequences

- **Automated homology modelling**
  - Swiss-model
  - Composer
  - Modeller
Last year (~1992) 91% of newly determined structures were similar to a previously available structure.
Recognizing Protein Folds

Interleukin 1β

Soybean Trypsin Inhibitor

Sequence identity = 9.5%

Mark Swindells
Advanced Threading Approaches

Prediction in 1990s and beyond

- Secondary Structure prediction
  - Major improvement with PHD and inclusion of homologous sequences

- Automated homology modelling
  - Swiss-model
  - Composer
  - Modeller

- Fold Recognition
- CASP
Assessment of Comparative Modelling in CASP2
Martin, MacArthur and Thornton
Proteins Suppl. 1:14–28, 1997


• 12 target structures, 9 were solved prior to the meeting: 8 by X-ray crystallography and 1 by NMR spectroscopy.

• Percentage sequence identity with the principal parent structure, which ranged from 20% up to 85%.

• 24 Groups submitted models
Conclusions Comparative Modelling CASP II

1. The most important factor in model accuracy is alignment accuracy. Below 25% SeqID, the alignments were so poor that the models were so far from reality to be not useful, even though the overall fold may be accurate.

2. The difference between the local and global RMSD for many loops (>60%) is large. The largest contribution to loop RMSD arises from these global shifts.

3. The all atom RMSDs correlate strongly with the Ca RMSDs. If the RMSD on C alphas is larger than 2 Å, the chi1 RMSDs are >70° (i.e. Approaching random).

4. Comparative assessment of the methods used is difficult, but correct inheritance from the parent is still probably the most important factor. Methods which have not achieved this adequately succeed less well.
Alignment Accuracy (1997) CASP II

![Alignment Accuracy Plot](image)
1997 Results

Observed accuracy (Ca RMSD, model vs. target) plotted against percentage sequence identity with the principal parent (SeqID).

CASP2 (open circles)
CASP1 (gray squares)
Ab initio CASP2 (black diamonds).
Why Predict Structure?

**Academic Challenge**
- Proves one’s understanding of the molecular forces which determine structure from sequence

**To learn about Function**
- What does it do?
  - Is it enzyme or binding protein?
  - If enzyme – what reaction does it catalyse?
  - Which ligands does it bind?
    - Small molecules
    - Polymers
    - What is structure of complex?
Protein Structure

The function of a protein may change during time, by duplication and subsequent evolution.

Molecular Function

From Structure to Function
Protein folds and functions

Haem-binding

Carbohydrate binding

DNA binding

ACR Martin et al, 1998 *Protein folds and functions* Structure 6, 875 - 884
Mapping the Proteome and the Metabolome
Triose phosphate isomerase

Highly promiscuous domain superfamily -
Shape Variance of Binding Pockets

**AMP**
- 1amu
- 1jp4
- 1qb8
- 8gpb
- 12as

**ATP**
- 1e2q
- 1b8a
- 1dy3
- 1gn8
- 1o9t

**NAD**
- 1mi3
- 1me1
- 1ib0
- 1rlz
- 1s7g
- 1s7g
How do EC numbers (ie enzyme function) relate to sequence/structure and vice versa?
Aspartate Amino Transferase Superfamily

- Aspartate Aminotransferase
- 2,2-Dialkylglycine Decarboxylase
- Tyrosine Phenolylase
- Ornithine Decarboxylase

Enzyme Classes:
- 2.6.1.1
- 4.1.1.64
- 4.1.1.17
- 4.1.99.2
Conservation of enzyme function in homologous superfamilies

- Total: 167 CATH superfamilies (by hand)
Evolving Enzymes

Changes in **enzyme function** during evolution are captured from sequence and structure based phylogenetic trees. Analysis of 276 enzyme families

# changes within same E.C. class = 2967 (89%)
  i.e. changes in substrates

# changes between E.C. classes = 360 (11%)
  i.e. changes in chemistry of reaction

From Molecules to Organisms

Protein/DNA → Structural Biology

Gene Expression → Macromolecular Complex

Cell Biology → Subcellular Structure

Developmental Biology → Cell

Organism → Computational Biology
Current Status of Protein Structure Prediction (by a current non-expert!)

- Fold recognition – HMMs dominate methods for recognition of distant relatives (sequence to sequence)
- Knowledge based modelling the norm – combining classic energy calculations with expectations from known structures and ‘other’ data
- Use of multi-servers is powerful – combining results of different methods
- Increase in compute power allows more in depth searching
- Homology modelling – fully automated and widely used
- Refinement and Docking still a challenge – both for small molecules and PPI – remains difficult since dependent on detailed packing and interactions
- More data, especially sequences – improves contacts prediction
Challenges 2012

• Calculating Free Energy ($\Delta G$) accurately
• Refinement
• Function Annotation for genomes using structure
  – Details of function
    • Biochemical function
    • Specificity
  – Prediction of biological function as well as biochemical function (test tube to \textit{in vivo})
• Application to Molecular Design
• From molecular to cellular structure – bridging scales/hybrid methods
Your Challenges 2012

• ?