EMBL-EBI PDBe PROTEIN DATA BANK EUROPE WORLDWIDE PDB PROTEIN DATA BANK

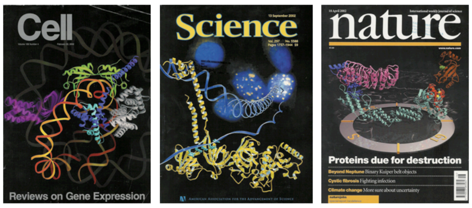
14 January, 2010 – Houston

Applied common sense

The why, what and how of validation
(and what EM can learn of X-ray)

Gerard J. Kleywegt
Protein Data Bank in Europe
EMBL-EBI, Cambridge, UK

Crystallography is great!!



- Crystallography can provide important biological insight and understanding

(and EM too, of course)

… but sometimes we get it (really) wrong

Nightmare before Christmas

Retraction

WE WISH TO RETRACT OUR RESEARCH ARTICLE "STRUCTURE OF MsbA from *E. coli*: A homolog of the multidrug resistance ATP binding cassette (ABC) transporters" and both of our Reports "Structure of the ABC transporter MsbA in complex with ADP•vanadate and lipopolysaccharide" and "X-ray structure of the EmrE multidrug transporter in complex with a substrate" (*1–3*).

The recently reported structure of Sav1866 (*4*) indicated that our MsbA structures (*1*, *2*, *5*) were incorrect in both the hand of the structure and the topology. Thus, our biological interpretations based on these inverted models for MsbA are invalid.

An in-house data reduction program introduced a change in sign for anomalous differences. This program, which was not part of a conventional data processing package, converted the anomalous pairs (I+ and I–) to (F– and F+), thereby introducing a sign change. As the diffraction data collected for each set of MsbA crystals and for the EmrE crystals were processed with the same program, the structures reported in (*1–3*, *5*, *6*) had the wrong hand.

The error in the topology of the original MsbA structure was a consequence of the low resolution of the data as well as breaks in the electron density for the connecting loop regions. Unfortunately, the use of the multicopy refinement procedure still allowed us to obtain reasonable refinement values for the wrong structures.

The Protein Data Bank (PDB) files 1JSQ, 1PF4, and 1Z2R for MsbA and 1S7B and 2F2M for EmrE have been moved to the archive of obsolete PDB entries. The MsbA and EmrE structures will be recalculated from the original data using the proper sign for the anomalous differences, and the new Cα coordinates and structure factors will be deposited.

We very sincerely regret the confusion that these papers have caused and, in particular, subsequent research efforts that were unproductive as a result of our original findings.

GEOFFREY CHANG, CHRISTOPHER B. ROTH,
CHRISTOPHER L. REYES, OWEN PORNILLOS,
YEN-JU CHEN, ANDY P. CHEN

Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037, USA.

References
1. G. Chang, C. B. Roth, *Science* **293**, 1793 (2001).
2. C. L. Reyes, G. Chang, *Science* **308**, 1028 (2005).
3. O. Pornillos, Y.-J. Chen, A. P. Chen, G. Chang, *Science* **310**, 1950 (2005).
4. R. J. Dawson, K. P. Locher, *Nature* **443**, 180 (2006).
5. G. Chang, *J. Mol. Biol.* **330**, 419 (2003).
6. C. Ma, G. Chang, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2852 (2004).

SCIENCE VOL 314 22 DECEMBER 2006 1875

(and EM too, of course)

The *why* of validation

- Crystallographers produce *models* of structures that *will* contain errors
  - High resolution AND skilled crystallographer ➔ *probably* nothing major
  - High resolution XOR skilled crystallographer ➔ *possibly* nothing major
  - NOT (High resolution OR skilled crystallographer) ➔ *pray* for nothing major

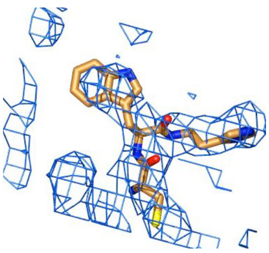
(and EM too, of course)

Why do we make errors?

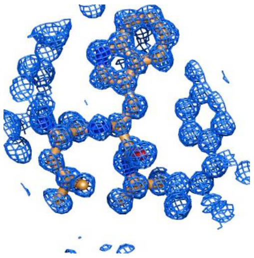
- Limitations to the data
  - Space- and time-averaged
    - Radiation damage, oxidation, … (sample heterogeneity)
    - Static and dynamic disorder (conformational het.)
    - Twinning, packing defects (crystallographic het.)
  - Quality
    - Measurement errors (weak, noisy data)
  - Quantity
    - Resolution, resolution, resolution (information content)
    - Completeness
  - Phases
    - Errors in experimental phases
    - Model bias in calculated phases

(and EM too, of course)

All resolutions are equal …



1ISR 4.0Å



1EA7 0.9Å

Why do we make errors?

- Subjectivity
  - Map interpretation
  - Model parameterisation
  - Refinement protocol
- Yet you are expected to produce a complete and accurate model
  - Boss
  - Colleagues
  - Editors, referees, readers
  - Users of your models
    - Fellow crystallographers, EM-ers, molecular biologists, modellers, medicinal chemists, enzymologists, cell biologists, biochemists, …, YOU!

(and EM too, of course)

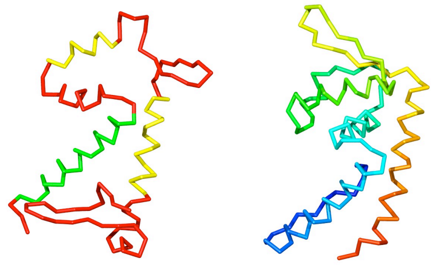
The *why* of validation

- Crystallographic models *will* contain errors
  - Crystallographers need to fix errors (if possible)
  - Users need to be aware of potentially problematic aspects of the model
- Validation is important
  - Is the model as a whole reliable?
  - How about the bits that are of particular interest?
    - Active-site residues
    - Interface residues
    - Ligand, inhibitor, co-factor, …

Great expectations

- Reasonable assumptions made by structure users
  - The protein structure is correct
  - They know what the ligand is
  - The modelled ligand was really there
  - They didn't miss anything important
  - The observed conformation is reliable
  - At high resolution we get all the answers
  - The H-bonding network is known
  - I can trust the waters
  - Crystallographers are good chemists
- In essence
  - We are skilled crystallographers and know what we are doing

The protein structure is correct?

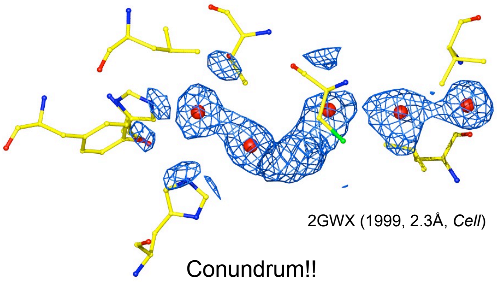


1FZN (2000, 2.55Å, *Nature*) 2FRH (2006, 2.6Å)

- One helix in register, two helices in place, rest wrong (*dread, mon*!)

- 1FZN obsolete, but complex with DNA still in PDB (1FZP)

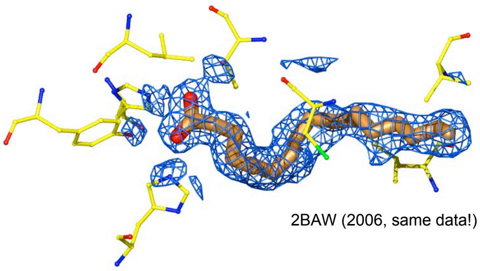
We didn't miss anything important?



2GWX (1999, 2.3Å, *Cell*)

Conundrum!!

Oh, *that* ligand!

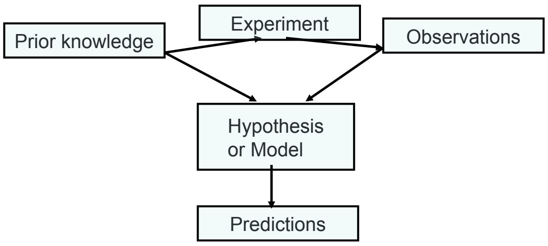


2BAW (2006, same data!)

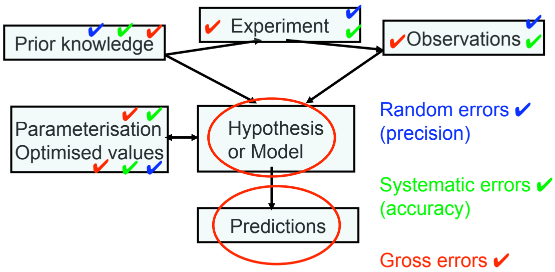
The *what* of validation

- Validation = establishing or checking the truth or accuracy of (something)
  - Theory
  - Hypothesis
  - Model
  - Assertion, claim, statement
- Integral part of scientific activity!

Science, errors & validation



Science, errors & validation



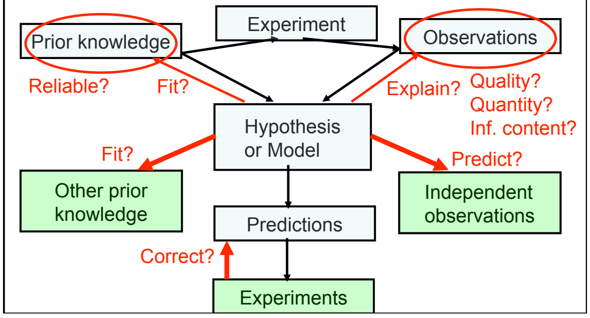
Random errors ✔ (precision)

Systematic errors ✔ (accuracy)

Gross errors ✔ (both)

Science not immune to Murphy's Law!

Science, errors & validation



The *how* of validation

- Q: What is a good model?
- A: A model that makes sense in every respect!

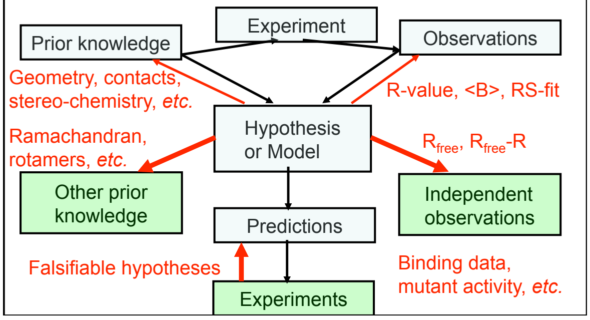
A good model makes sense

- Chemical
  - Bond lengths, angles, chirality, planarity
- Physical
  - No bad contacts/overlaps (incl. implicit H-atoms), close packing, reasonable pattern of variation of Bs, charge interactions
- Crystallographic
  - Adequately explains/predicts experimental data (R, R_{free}, R_{free} - R), residues fit the density well

A good model makes sense

- Protein structural science
  - Ramachandran, peptide flips, rotamers, salt links, prolines, glycines, buried charges, residues are “happy” in their environment, hydrophobic residues in core
  - Comparison to related models
- Statistical
  - Best hypothesis to explain the data with minimal over-fitting (or “under-modelling”!)
- Biological
  - Explains observations (activity, mutants, inhibitors)
  - Predicts (falsifiable hypotheses)

Science, errors & validation


Prior knowledge

Experiment

Observations

Geometry, contacts, stereo-chemistry, *etc.*

R-value, <B>, RS-fit

Hypothesis or Model

Ramachandran, rotamers, *etc.*

R_{free}, R_{free}-R

Other prior knowledge

Predictions

Independent observations

Falsifiable hypotheses

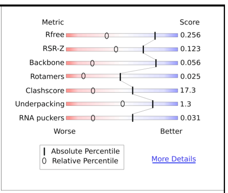
Binding data, mutant activity, *etc.*

Experiments

Validation in a nutshell!

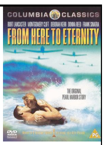
- Compare your model to the experimental data and to the prior knowledge. It should:
  - ***Reproduce*** knowledge/information/data used in the construction of the model
    - R, RMSD bond lengths, chirality, …
  - ***Predict*** knowledge/information/data ***not*** used in the construction of the model
    - R_{free}, Ramachandran plot, packing quality, …
  - Global and local
  - … and if your model fails to do this, there had better be a plausible explanation!

X-ray VTF



- Validation pipeline
  - State-of-the-art methods
    - Phenix, WhatCheck, MolProbity, EDS,…
  - Will produce a report (PDF)
    - Can be submitted to journals
    - Mandatory in the future? (IUCr, PNAS)

Where to go from here?



- Download and read:
  - GJ Kleywegt. Validation of protein crystal structures. *Acta Crystallographica* **D56**, 249-265 (2000) (and many references therein)
  - GJ Kleywegt. On vital aid: the why, what and how of validation. *Acta Crystallographica*, **D65**, 134-139 (2009)
- Do this web-based tutorial:
  - http://xray.bmc.uu.se/embo2001/modval

勿因一时疏忽
破坏永恒美好
A SINGLE ACT OF CARELESSNESS LEADS
TO THE ETERNAL LOSS OF BEAUTY