

Stereochemistry and *ab initio* phasing

Gerard Bricogne
Global Phasing Ltd.
Cambridge, UK

Ab initio phasing

- The notion comes from small-molecule crystallography: use only amplitude data for symmetry and Friedel unique reflections, and chemical composition:
 - no anomalous scattering data;
 - no structural knowledge.
- How relevant is it to today's macromolecular crystallography?
- Is this definition too restrictive, too academic?

However, keep “*ab initio*” special ...

trial solutions. Attention is here focused onto the case in which diffraction data of one isomorphous derivative are additionally available. It is shown that in such a case direct *ab initio* solution of protein structures is feasible. Tests based on calculated diffraction data suggest the procedure to follow for a possible success.

A39, 685–692], is reconsidered. Experimental tests show that the formula is potentially able to estimate phases accurately provided 30–40% of the electron density is correctly located. The formula may be used for refining the phase values obtained by isomorphous derivative techniques as well as for extending the phasing process to a resolution higher than the derivative resolution.

Criteria for the solution of the phase problem

- The one and only criterion for the correct phases is the chemical validity of the structure indicated by the electron density, and its compatibility with prior chemical knowledge (unless it corrects it!).
- Paul Ewald, in *Fifty Years of X-ray diffraction*:
 - “when the phases are good, the map looks like a nice dish of *fried eggs*;
 - when the phases are bad, we have *scrambled eggs*.”
- We need *a mathematical criterion for fried eggs* expressed in terms of amplitudes and phases.

Amazingly, such a criterion exists at atomic resolution: Sayre's equation

The Sayre equation is the reciprocal-space equivalent of a simple relationship in real space, $\rho(r) \propto \rho^2(r)$, which is valid for sharply peaked electron-density maps (the so-called atomicity condition). The Sayre equation reads (Sayre, 1952)

$$\mathbf{F}(\mathbf{h}) = g(\mathbf{h}) \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h} - \mathbf{k}), \quad (1)$$

where $g(\mathbf{h})$ is a resolution-dependent form factor.

A weaker criterion: positivity

Karle & Hauptman (1950), Toeplitz (1911)*,
Caratheodory (1911)*:

The positivity of the electron density function implies an infinite family of inequalities for determinants of convolution matrices formed from structure factors.

(*) both papers were published back-to-back in the same issue of *Rend. Circ. Mat. Palermo*.

Probabilistic direct methods

- Triple phase relationship (Cochran, Hauptman & Karle, ~1953).
- Tangent formula (Karle & Hauptman, 1956):
- Advent of multisolution methods and of software for automated structure solution (MULTAN, 1968ff).

$$\tan(\varphi_{\mathbf{H}}) = \frac{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{\mathbf{H}-\mathbf{K}}| \sin(\varphi_{\mathbf{K}} + \varphi_{\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} |E_{\mathbf{K}} E_{\mathbf{H}-\mathbf{K}}| \cos(\varphi_{\mathbf{K}} + \varphi_{\mathbf{H}-\mathbf{K}})},$$

Multiresolution (multi-trial) strategies

- Analyse the connectivity of the phase relationships.
- Make whatever phase choices are necessary to fix the origin and define the enantiomorph.
- Sample phase values for a “basis set” of influential phases (e.g. 128 values of quadrant phase choices).
- Propagate these values by the tangent formula.
- Apply various “figures of merit” to select the most likely phase sets among the 128.
- Examine maps by peak picking, apply known rules of chemistry (N.B.: *after* phasing).

Use of quartets

- Besides triplets $\Phi_h + \Phi_k + \Phi_{-h-k}$ one can consider quartets $\Phi_h + \Phi_k + \Phi_l + \Phi_{-h-k-l}$.
- Schenk (1974) showed empirically, and Hauptman (1975) theoretically, that looking at the amplitudes of the “cross-terms” at $h+k$, $h+l$, $h+m$ (where $m=-h-k-l$) could indicate negative as well as positive values for the cosine of the quartet.
- These negative quartets were useful new figures of merit, especially to avoid the “uranium atom” solution.

The Golden Age of Probabilistic Direct Methods

- Equipped with triplets, quartets, tangent formula expansion and fragment recycling, probabilistic direct methods were able to solve routinely light-atom structures with 100-150 atoms (MULTAN, SHELX).
- This plateau was due to the weakening of TPRs, or equivalently by the rarefaction of large E's with increasing structure size, and by lack of good figures of merit by which to rank large numbers of trials.
- Random-start strategies with large basis sets helped alleviate this problem to some extent (RANTAN).

Intermezzo (all Erice's fault ...)

A fresh look at the probability basis of direct methods (Bricogne, 1984):

- Edgeworth's series can be improved by the saddlepoint approximation;
- the latter bears an intriguing relationship to the Maximum-Entropy Method;
- this approach emphasises the progressive build-up of non-uniformity in the prior probability distribution of randomly placed atoms (assumed uniform in direct methods);
- a constructive procedure was obtained to build multivariate joint distributions of any collection of structure factors, to the Gaussian level of approximation or beyond;
- a multivariate generalisation of the TPR was shown to be related to *maximum-entropy extrapolation*;
- a multivariate generalisation of quartet figure of merit was shown to be related to a *log-likelihood gain in the "second neighbourhood"* of the basis set (defined as a product of Rice distributions – called by their proper name for the first time);
- Led to a multisolution strategy by tree-searching, driven by maximum-entropy extrapolation and pruned by log-likelihood gain in $N_2(H)$ – i.e. direct methods in a new form.

... and its aftermath

- Six papers with Chris Gilmore in 1990-93 on implementing this new approach of “Entropy Maximisation and Likelihood Ranking” and applying it to
 - small molecules,
 - powders,
 - a small protein (APP),
 - electron crystallography,
 - rescuing a difficult protein phasing situation (with Charles Carter).
- Demonstrated the power of likelihood as a figure of merit for hypothesis-ranking, and hence as a refinement criterion.
- This use of likelihood was also cited in the rationale for R^{free} .
- Entropy maximisation has remained difficult to harness in a reliable manner, in spite of its unquestionable power.

Meanwhile, in the main stream

- Herb Hauptman returns to real space:
 - The impasse at the stage of “getting off the ground” caused by the shortage of large E’s is overcome by a random-start strategy *in real space* (random structure instead of random phases: fried eggs by design, in search of positions).
 - Initial phases from a random starting structure are expanded by the tangent formula, then improved by use of a “minimal function” derived from old-fashioned probability theory.
 - A new (hopefully less random) structure is then obtained by peak picking out of a Fourier map.
 - The process is iterated in a method called “Shake and Bake”.

- ... and George Sheldrick finds a shortcut:
 - Instead of using a minimum function to improve phases, the “Half-Baked” procedure (now SHELXD) uses “iterative peaklist optimisation” or “random-omit maps”, and a correlation coefficient on I’s as a figure of merit.
- These dual-space methods add to the long tradition of algorithms alternating between real and reciprocal space (Barrett & Zwick 1972; Gerchberg & Saxton 1975; Bricogne 1974,76; Wang, 1985; ...).
- They pushed the limits of *ab initio* phasing firmly into the macromolecular range, breaking the 1000-atom barrier here (in Erice) in 1997.

Sheldrick's rule

- In 1990 George Sheldrick wrote:

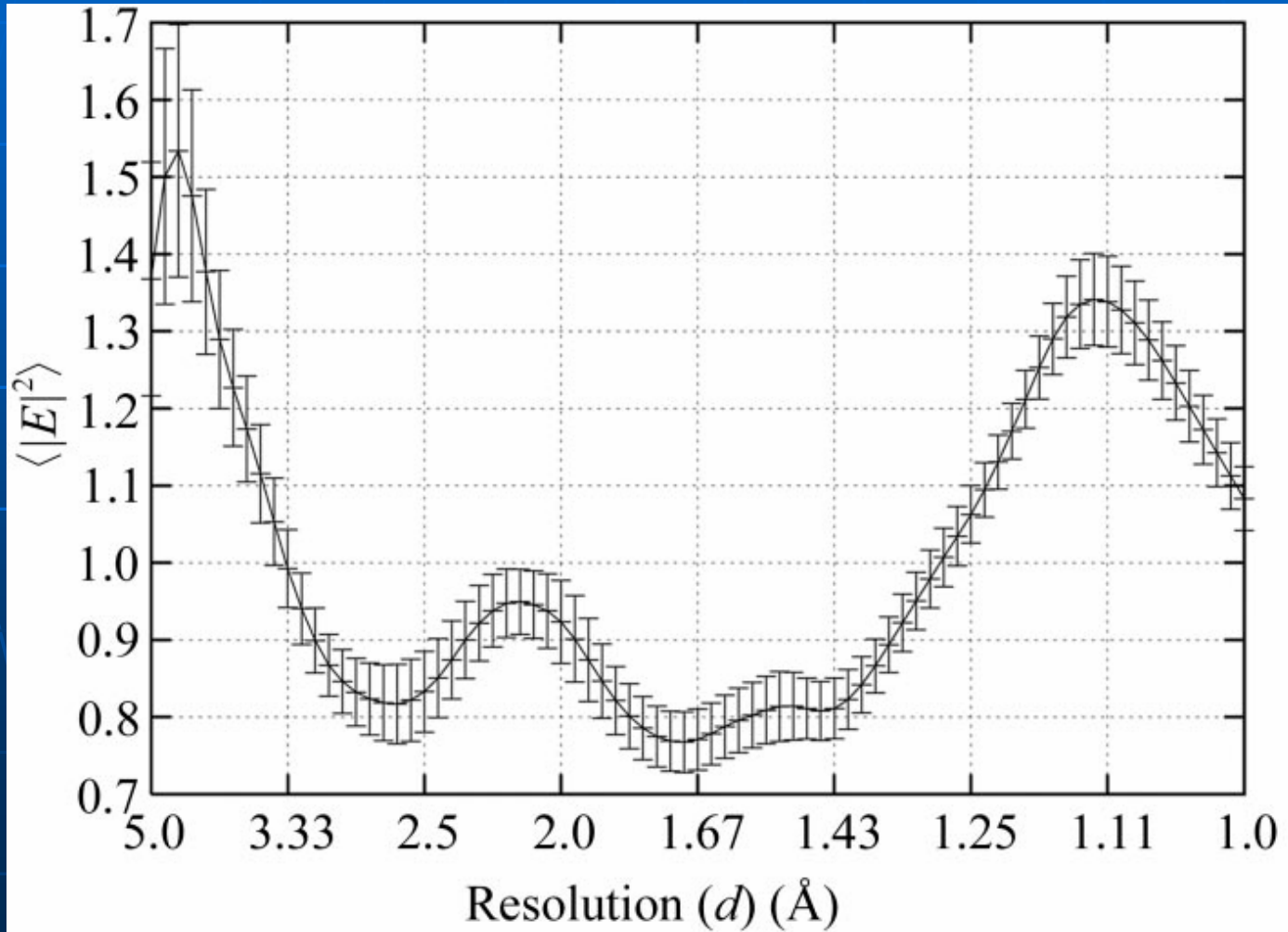
“Experience with a large number of structures has led us to formulate the empirical rule that if fewer than half the number of theoretically measurable reflections in the range of 1.1 to 1.2Å are ‘observed’ [i.e. have $F > 4\sigma(F)$], it is very unlikely that the structure can be solved by direct methods ... This rule simply reflects the assumption of resolved atoms, which is often invoked in direct methods.”

- It is remarkable that, although the *size* of structures that are now solved by dual-space direct methods has increased dramatically, *this rule has held firm* – with exceptions only when there are heavy atoms to make phasing easier.
- Essentially the same limitation seems to apply asymptotically to other semi *ab initio* phasing methods (e.g. ACORN) when the amount of “seed” information decreases.

Missing information

- Atomic resolution data provides enough information to solve a structure without any prior chemical knowledge: indeed, our knowledge of stereochemistry was derived from high-resolution X-ray structures.
- Lowering resolution decreases the amplitude information available.
- Missing information can only be replaced by prior knowledge or by hypotheses.
- At some critical stages, prior chemical knowledge will have to be brought in to compensate the loss of data accompanying the loss of resolution.
- Sheldrick's rule is an instance of such a phenomenon.

Averaged $|E|^2$ profile from 700 protein structures (Morris & Bricogne, 2003)



Debye's formula, and beats between distances

Atom	Atom	Distance (Å)	Difference (Å)
C^i	O^i	1.23	1.15
C_{α}^i	N^{i+1}	2.38	
C^i	N^i	1.32	1.13
C_{α}^i	C^{i+1}	2.45	
C_{α}^i	C^i	1.52	1.08–1.18
$C_{\alpha,\beta}^i$	$C_{\gamma,\delta}^i$	2.6–2.7	
C_{α}^i	C_{α}^{i+1}	3.81	1.11–1.21

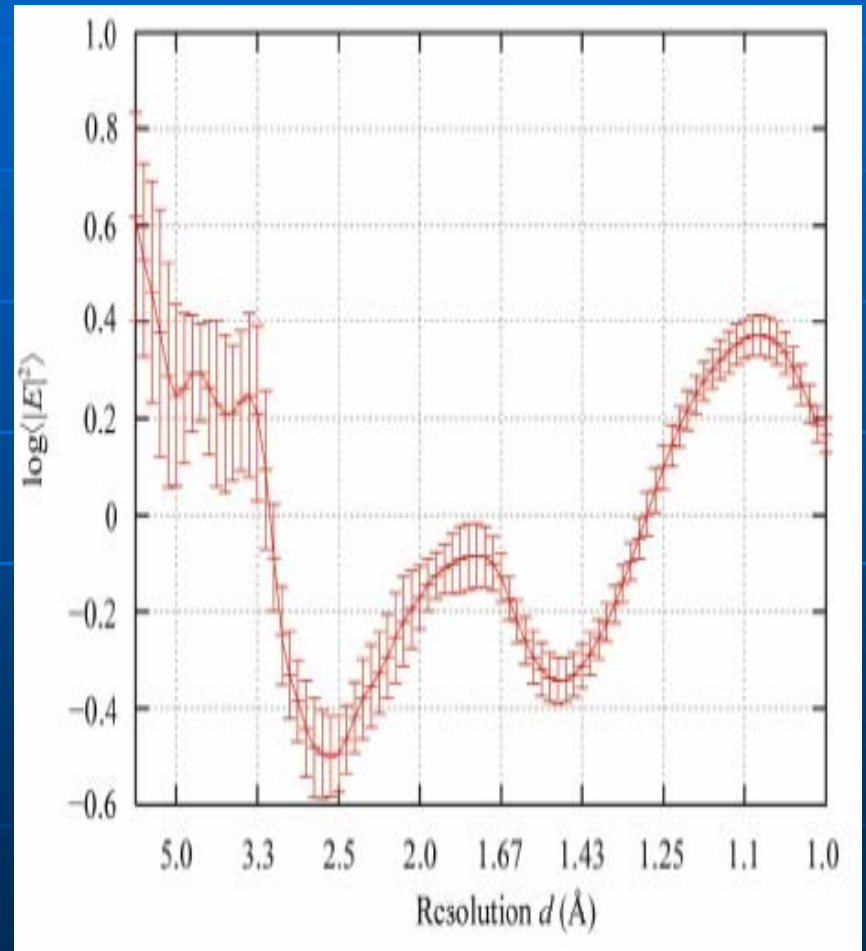
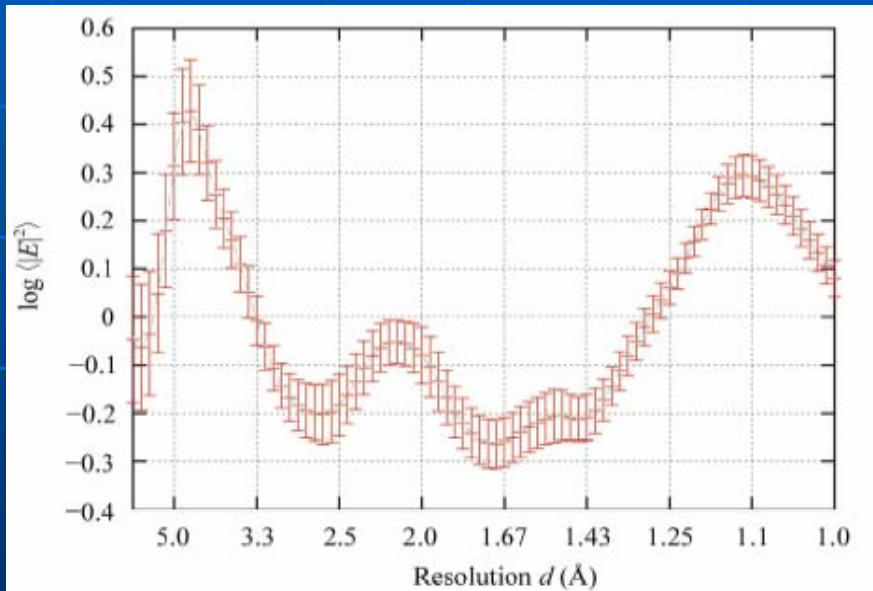
$$I(d^*) = \sum_i f_i^2(d^*) + \sum_i \sum_{i \neq j} f_i(d^*) f_j(d^*) \text{sinc}(2\pi d^* r^{ij}),$$

A structural basis for Sheldrick's rule

- Contrary to initial rationalisations (Dauter, Lamzin & Wilson, 1997), Sheldrick's rule is not only a case of an observations-to-parameters ratio.
- The amplitude data above 1.2Å indicates regularities in distance distributions which follow from specific bond angles, not only bond lengths.
- The “stereochemistry-free” direct methods are unable to recover from the loss of this information – no matter whether we consider classical or dual-space direct methods.

Ubiquity of the 1.1Å peak

(Morris, Blanc & Bricogne, 2004)



Replacing data by prior knowledge

- Information not given by the X-ray data must come from prior knowledge, or else a huge dilution of correct solutions among incorrect ones will result (“needle in haystack”).
- Any step where atoms were previously treated as statistically independent must be made “stereochemically aware”.
- For biological macromolecules, there is the PDB and there are evolutionary relationships.

Extending the classical path: the Micromolecular Replacement (μ MR) Method.

- Build joint probability distributions of structure factors on the basis of randomly located and oriented fragments, instead of randomly located atoms (Bricogne 1994, 95, 97).
- The ME formalism extends naturally to such a model, giving “stereochemically aware” structure factor statistics and likelihood functions.
- The choice of the composition of the “soup” of fragments gives an extra dimension to the search, allowing a recourse to the hierarchical structure of proteins.
- Fold classification, structural templates, and other finer-grained systematics of protein structure can be invoked to help choose the initial composition of the soup and adjust it as phase hypotheses are formed, evaluated, and expanded.
- This is still based on a search over phases of strong reflections.

Extending the dual-space path: stereochemically-aware baking.

- Replace the initial random structure of the current dual-space methods by a hypothetical macromolecular structure, or at least by a structure which is locally macromolecule-like.
- Try to enrich for macromolecular features at each real-space “baking” stage, by detecting and connecting valid structure fragments.
- Isabel Uson and George Sheldrick are working along these lines – any results on the way?

Automated formation of structural hypotheses: the ARP/wARP paradigm

- ARP/wARP is a dual-space algorithm which works from some starting phase information:
 1. it places free atoms in an electron-density map;
 2. it identifies pieces of main-chain (from $C\alpha$ - $C\alpha$ distances and the Ramachandran plot) and generates the corresponding stereochemical restraints;
 3. at a later stage, it may dock side-chains and include them in the restrained model;
 4. it refines the free and restrained atoms (the “hybrid model”) with REFMAC, producing a 2Fo-Fc map;
 5. then iterates back to (1).
- Efficiency comes from (2), which is very specific and cuts through a lot of inconclusive searching.
- The algorithm can not only interpret maps, but can also do extensive rebuilding on poor-quality MR models.

Combinatorics of uncertain restraints

- The real-space counterpart of phasing is the assignment of specific stereochemical restraints to subsets of atoms.
- ARP/wARP does so by looking for unambiguous local interpretations of an electron density map and extending them thanks to the phase (hence map) improvements produced by the restraints.
- To move towards ab initio phasing, we need to search through ambiguous local interpretations – here arises the “haystack”.
- This has been attempted by Scheres & Gros under the name of Conditional Dynamics.

Conditional Dynamics (1)

- Described as “incorporating extensive geometric prior information without the necessity of an explicit interpretation of the electron-density map”.
- Considers all possible assignments of chemical types to connected clusters of atoms close in space, and all possible local interpretations of each cluster in terms of known protein structure elements.
- The uses a target function which will be minimal for a cluster of atoms with correct chemical identities in an expected local conformation. Similar to the knowledge-based interaction functions of M. Sippl.

Conditional Dynamics (2)

- 2001: built helices of a poly-Ala 4-helix bundle from randomly distributed atoms, “in a simple artificial test case” from (calculated) 2A data.
- 2003: three protein structures with large randomly distributed coordinate errors were refined by CD. Good results on the all-alpha structure, did less well on mixed or all-beta.

Conditional Dynamics (3)

- 2004: Automated model building on three structures, at 2.4, 2.6 and 3.0Å resolution.
 - CNS refinement with phase information and CD force field instead of standard geometric restraints, starting from free atoms placed in density.
 - Same, but with multiple random-atom start, again using phase information in the *X*-ray term of the refinement objective function.
 - Result: worse than ARP/wARP, better than RESOLVE, but very slow.

Comparison of results in automated building mode

Statistics of automated model building of three crystal structures at medium to low resolution.

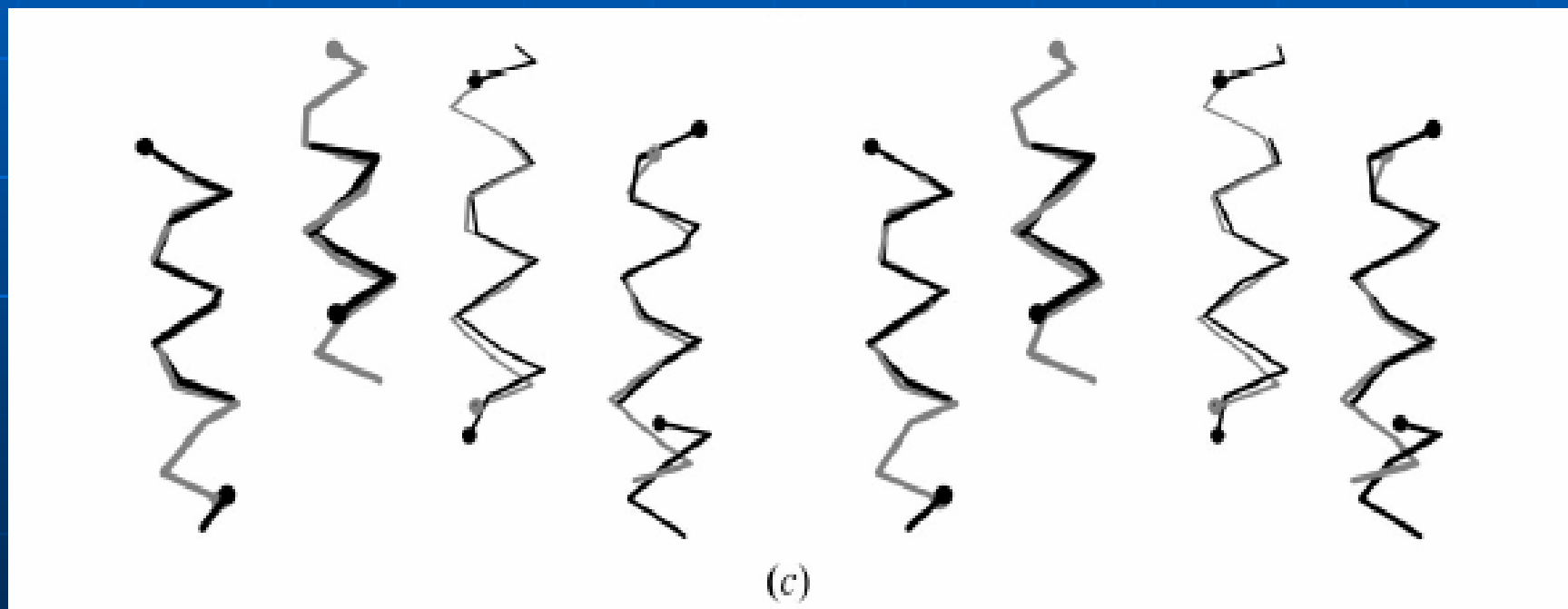
	vWF-A3 (2.4 Å)			NspA (2.6 Å)			LAPP (3.0 Å)		
	CO	ARP	RESOLVE	CO	ARP	RESOLVE	CO	ARP	RESOLVE
No. residues built	148	160	170	121	130	101	169	232	136
Fraction built (%)	80	87	93	78	84	65	64	87	51
No. chains	20	8	4	16	6	10	38	11	13
R.m.s.d.† (Å)	1.5	1.9	0.9	1.3	1.8	0.9	1.2	1.3	1.8
$\langle \Delta\varphi \rangle \ddagger$ (°)	27.2	36.0	23.9 (22.7)§	35.8	33.6	42.4 (35.6)	25.4	27.9	56.5 (26.0)
$\langle \cos(\Delta\varphi) \rangle \P$	0.67	0.56	0.72 (0.69)§	0.53	0.60	0.46 (0.52)	0.67	0.64	0.32 (0.66)
CPU (h)	36	1	15	38	2.5	18	105	1.5	14

† Root-mean-square coordinate deviations, or coordinate errors, were calculated based on the distance between atoms in modelled protein fragments to the nearest atom with a corresponding atom label in the refined structure. ‡ Amplitude-weighted mean phase error calculated with respect to the refined structures. § For *RESOLVE*, the phase errors of the resulting electron-density maps are given in parentheses. ¶ Unweighted mean cosine of the phase error with respect to the refined structures. For calculation of both amplitude-weighted and unweighted phase errors all atoms of the resulting models were taken into account, *i.e.* for models generated by conditional optimization (CO) or *ARP*/*wARP* (ARP) atoms that were not recognized as part of a protein fragment were also included.

Conditional Dynamics (4)

- 2004: Ab initio phasing from real 2A data of the 4-helix bundle used in 2001.
 - Multiple starts without phases;
 - then targeting towards weighted mean phases for a cluster of 17 models.
 - Some hand-tweaking necessary.
 - Final phase error 76 degrees.
 - Final map correlation coefficient 0.37.
 - Huge CPU cost, but does look intriguing.

Stereoview of the C α trace of the model with the highest σ_A value



Top-down vs. Bottom-up.

- Conditional dynamics and dual-space methods with fragment search may be described as “bottom-up” strategies, as they (would) go from atoms towards larger fragments.
- The use of fold templates, and Combinatorial MR, on the contrary, are “Top-down” approaches, from overall structure towards fine detail.

Time for a paradigm shift?

As the emphasis moves from *X-ray* data towards structural knowledge as the primary source of information, *ab initio* phasing becomes more like ...

... *Structure Prediction with X-ray data!*

Proposal

Along with CASP, organise

CASP-X

where the crystallographers would make available (say) their 5.0Å resolution data.

Summary

- *Ab initio* macromolecular phasing may seem to have some degree of mythical (“Holy Grail”) and artificial character, when anomalous scattering is becoming so powerful (however, think of eukaryotic proteins).
- Any progress in *ab initio* phasing is important through its impact on substructure solution.
- *Ab initio* phasing is the nexus where all hard problems which occur elsewhere in macromolecular crystallography are seen in their most essential form (especially: testing ground for statistical concepts and methods).
- There are many exciting prospects ahead – ***young blood needed!***

Acknowledgements

All the past Erice Schools on Direct Methods
(R.I.P.)