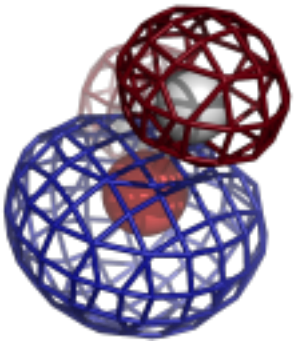
VOLUME TEN



COMPUTATIONAL CRYSTALLOGRAPHY NEWSLETTER

JANUARY MMXIX

MARCO, MAP-TO-MODEL V2

Table of Contents

Editor
Nigel W. Moriarty, NWMoriarty@LBL.Gov

Phenix News

Announcements

Workshop at the meeting of the American Crystallographic Association, Northern Kentucky Convention Center, Saturday, July 20, 2019

A workshop will be held at the next ACA annual meeting in Kentucky. The title of the all-day program – "Introduction to PHENIX for Electron Cryo-Microscopists" – indicates the target audience. Updates to schedules and the cost are available from the ACA homepage. The course is limited to 50 participants so book early.

Expert advice

Fitting Tip #17 – Asn and Gln are remarkably different

Jane Richardson, David Richardson and Christopher Williams, Duke University

Expectations of similarity

With the same amide functional group and only one carbon difference in sidechain length, Asn and Gln are usually considered one of the most similar pairs of amino acids. Looking at their 2D schematic diagrams (figure 1) or at their chemical makeup seems to confirm that idea, also reinforced by classic lists of conservative amino-acid replacements. But if one looks at what they each can or can't

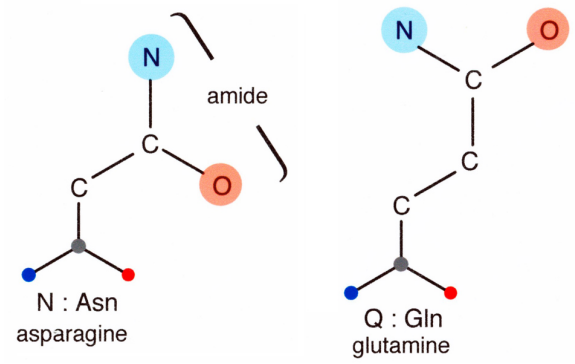


Figure 1: Schematics of Asn and Gln amino acids.

do in the context of protein 3D structures, that one-bond difference makes a huge change to their capabilities and personalities.

Multi-dimensional ϕ,ψ,χ plots

Asn has only two degrees of freedom, and has both donor and acceptor groups very close to the backbone, where they can form many distinct sidechain-backbone hydrogen bonds. This leads to a number of tight and unusual clusters in the multi-dimensional **ϕ,ψ,χ** space. Glutamine has more degrees of freedom but awkward constraints from the extra tetrahedral group and can actually H-bond back to the mainchain in only a few of its possible conformations.

As one way of showing those differences, figure 2 compares a diagonal view for the most informative 3D projections of the **ϕ,ψ,χ** plots for Asn and Gln. It uses data from the Top8000 database at 70% sequence identity (from the RCSB PDB clusters), quality-filtered at both chain and residue levels, including amide flips (Hintze, 2016). There are about 54,000 Asn and 37,000 Gln residues.

Each panel of figure 2 shows a 3D plot of **ϕ,ψ**, and **χ2** for Asn or of **ϕ,ψ**, and **χ3** for Gln, as divided down the vertical columns by the three **m,p,t** bins of **χ1** for Asn or of **χ2** for Gln. The viewpoint is rotated about 45° left from a pure Ramachandran-plot **ϕ,ψ** projection, to enable spotting 4D clusters vs spreads of the terminal amide orientations. Colored stars mark positions of local structure motifs discussed here.

From figure 2, the simplest overall observation is that the datapoint distribution for Asn is much more diverse and complex than that for Gln. The main reason is that the Gln amides are farther from the backbone so their orientations can spread freely across more of their range. The second overall observation is that datapoints are quite dense within 90° of zero and absent or sparse within 90° of 180° (Lovell 1999). Near 180° the NH2 group clashes with backbone or Cβ hydrogen atoms.

Glutamine characteristics

Especially when Gln **χ2** is trans, the **χ3** distributions are broad smears across their allowed range away from 180°. There are some exceptions where a small cluster appears at 180° offset from the central **χ3** maximum, usually seen at far left on the panels in figure 2. This data has been amide-flip corrected by MolProbity's H addition process that uses both H-bonding and clashes in the context of entire H-bond networks and seldom declares flips incorrectly (Word 1999). But it has a threshold of score difference below which it will keep the original orientation; that means there will still be some incorrect flip states remaining, which account for most points in the 180°-translated, fainter (~10%) patterns seen in figure 2. The same thing happens for Asn. After our realization that such cases are rather frequent, we plan to add a prior-probability term to the flip-correction process.

Glutamine does have some preferred patterns of amide-to-backbone H-bonds, but they nearly all occur in otherwise-favorable rotamers and so do not form tight, well-separated clusters. An especially notable Gln motif forms the helix "cap box" H-bond from the Gln OE1 to the backbone NH of the N-cap residue that starts an α-helix, as shown in figure 3 for a helix in λ repressor. This case has a Thr N-cap, although Asn is much commoner, as described below. The Gln cap-box conformation happens to work in exactly the commonest of all Gln rotamers (**mt**-30), so it just accounts for a modest part of the center of that strong, elongated cluster (hotpink * in the Gln **t** panel of figure 2).

Asparagine characteristics

Asparagine, in contrast to Gln, has many isolated clusters in unusual positions that represent distinct local structural motifs, most with specific sidechain-backbone H-bonding. Figure 4 shows the motifs for two distinct datapoint clusters of Asn sidechains on

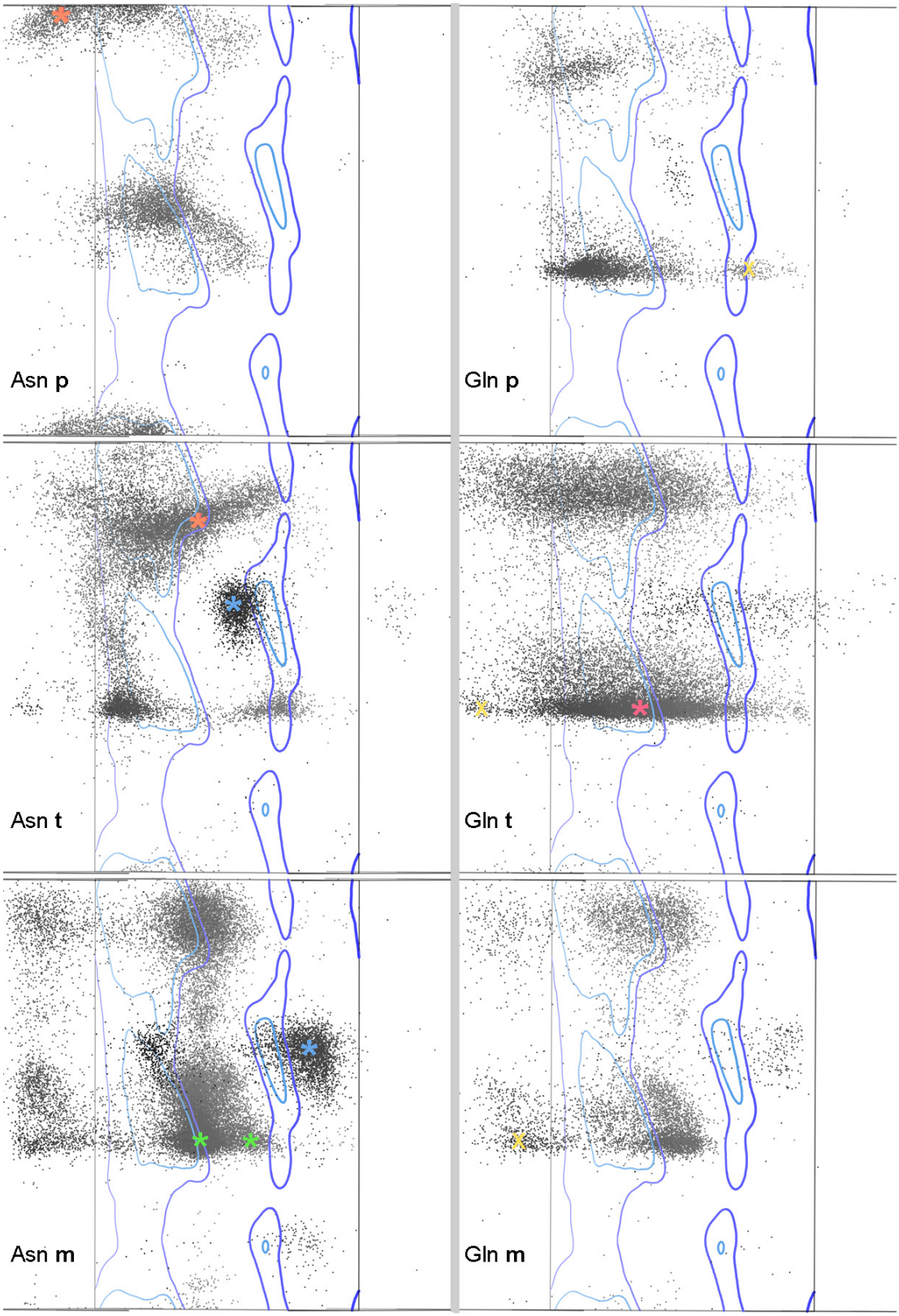


Figure 2: Diagonal views into the multidimensional ϕ,ψ,χ datapoint distributions for Asn, grouped by by $\chi 1$ bins, and for Gln, grouped by $\chi 2$ bins and labeled as **p, t, m** for ~60°, ~180°, and ~ -60°. View is turned left by 45° from straight-on ϕ,ψ, and stars mark local structural motifs discussed in the text.

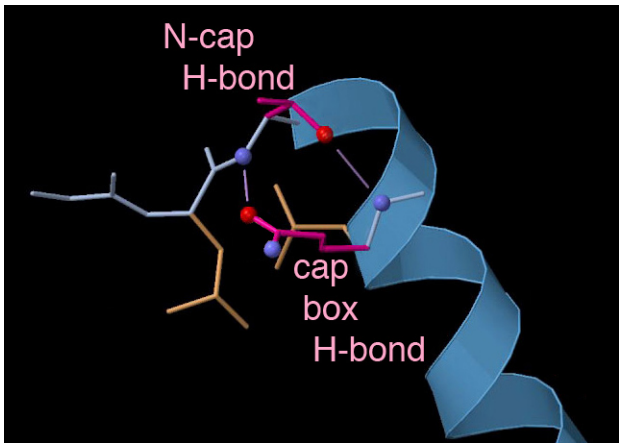


Figure 3: A helical "cap-box" Gln

regular α-helix (a relatively rare location for Asn), clearly separated in χ2 angle by an energy barrier. At left, the α**m**-80 rotamer enables the Asn NH2 to H-bond with the i-4 CO of the preceding helix turn, opening the backbone a bit, but still allowing the normal helical H-bonds as well. At right, the much more common α**m**120 rotamer places the entire sidechain in good vdW contact with the outer surface of the preceding turn of regular helix. That conformation also places the Asn OE1 in its favored position in vdW contact with the following peptide (Lovell 1999). The peaks for those rotamers are marked with green stars in the Asn **m** panel of figure 2.

Many of the local Asn motifs are enabled by the odd fact that the Asn sidechain is a very good mimic for a residue-unit of backbone, as illustrated in figure 5. At top is a short piece of extended backbone, with a sidechain (blue) going down and back. Below, on the left half, the main chain and side chain switch places; the backbone now goes down and back and the Asn sidechain (blue) mimics extended backbone with χ2 = 0° and is set up to make the previous CO H-bond equivalently (Richardson 1989). This similarity of the Asn sidechain to backbone (looking in the N-terminal direction from a given Cα) is also probably what lets Asn be the only non-Gly residue good at adopting +ϕ backbone conformations – the Cα of an Asn has two nearly identical substituents and thus is only

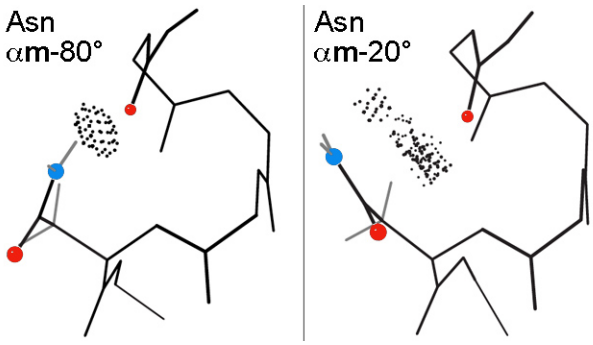


Figure 4: Two distinct conformations of Asn on α-helix

weakly asymmetric, so that normal Rα is not much better than Lα. The strong Lα peaks for Asn are marked with blue stars in the Asn **t** and **m** panels of figure 2, each with one rotamer. Asn with χ1 = **p** cannot adopt either Rα or Lα (see figure 2)

There are many local motifs in which the Asn sidechain mimics backbone. One such case is a "pseudo-turn", where the Asn takes the place of the first peptide in a tight turn, using Asn's most common local H-bond to backbone: Oδ1 to the i+2 backbone NH (Richardson 1981).

An important second case of such mimicry is at helix N-caps, where Asn is especially good at competing with backbone for its usual i+4 helical H-bond. As shown in figure 6, there are two possible rotamers that can make that

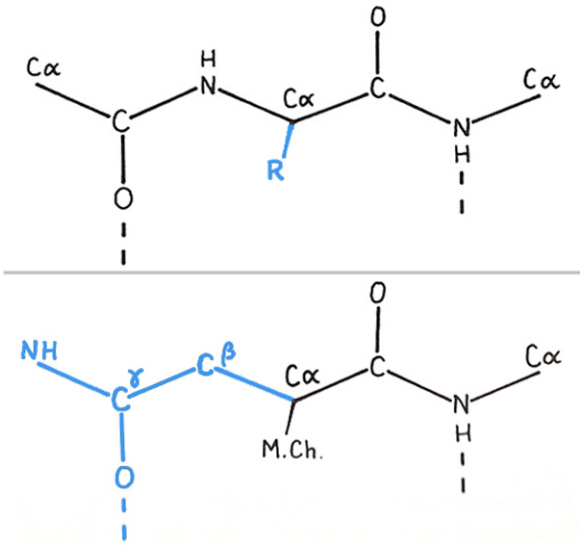


Figure 5: How an Asn sidechain mimics backbone.

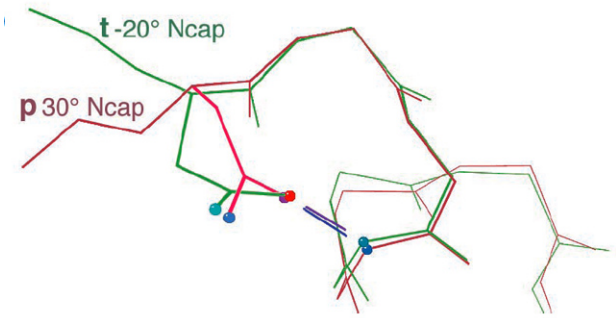


Figure 6: Two versions of Asn N-cap on α-helix.

H-bond, but **p**30 is the more common, since it is a closer geometric mimic and puts the displaced backbone in β rather than polyproline II conformation. Asn is not only the commonest N-cap residue, it is by a large factor the most specific, very strongly preferring the N-cap position and disfavoring the surrounding N-1 or N+1 to 3 positions (Richardson 1988). The sequence placement of an Asn, then, exerts a strong influence on where a helix starts, and on the direction from which the chain enters.

Yet another Asn backbone-mimic motif is to provide one more H-bond (sidechain-backbone) past β-sheet backbone H-bonding between two β-strands, either parallel or antiparallel. Usually only one such H-bond is formed, but figure 7 shows a case where the Asn amide forms two H-bonds to the opposite β-strand and a third to a separate part of the chain.

The bottom line

Glutamine is rather a "plain vanilla" sidechain, with Ramachandran plot and positional preference closest to the average of all residues. Asparagine, in contrast, has very distinct and opinionated conformational possibilities, both because it has H-bond donors and acceptors close to the backbone and because it can mimic a backbone residue.

When modeling Asn or Gln residues into a density map or evaluating them later, if you

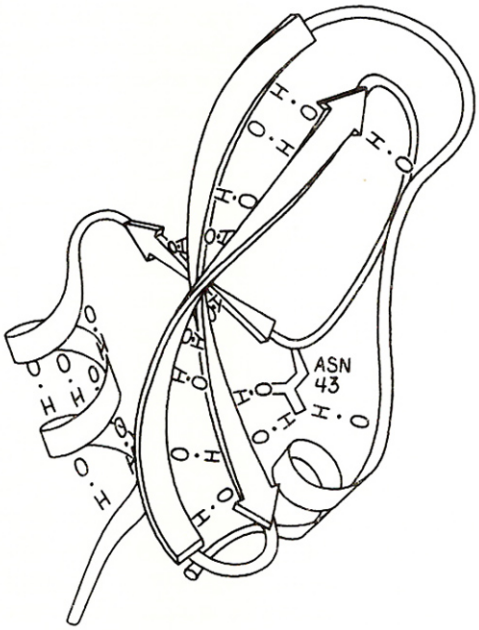


Figure 6: An extra Asn H-bond past the end of β-sheet.

have assigned a conformation with the final χ angle closer to 180° than to 0°, try the flipped alternative. If the orientation closer to 0° looks at least nearly as good by other criteria, then use that (more probable) alternative.

When modeling Asn sidechains, look for approximation to one of its distinctive local motifs such as N-caps, pseudo-turns, Lα backbone, interactions with the previous helix turn, H-bonding across the end of a β-strand pair, etc. If your Asn and its neighborhood are close to the arrangement of a typical Asn motif, try restraining the appropriate rotamer and H-bonds. In loops, look for any plausible sidechain-backbone or sidechain-sidechain H-bonding opportunities accessible with small changes.

If thinking about mutations or evolutionary relationships, don't consider Asn and Gln as conservative replacements for each other unless you know their function is either complete solvent exposure or very long-range amide H-bonding. Asn is more often the best replacement for a Gly than for a Gln.

References:

Richardson JS, Richardson DC (1988) Amino Acid Preferences for Specific Locations at the Ends of Alpha Helices, *Science* **240**: 1648-1652

Richardson JS, Richardson DC (1989) Principles and Patterns of Protein Conformation, chapter 1 in Prediction of Protein Structure and the Principles of Protein Conformation, ed. G. Fasman, Plenum Press, 1-98

Lovell SC, J.M. Word, Richardson JS, Richardson DC (1999) Asparagine and Glutamine Rotamers: B -Factor Cutoff and Correction of Amide Flips Yield Distinct Clustering, *Proc. Natl. Acad. Sci. USA*, **96**: 400-405

Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation, *J Mol Biol*, **285**: 1735-1747

Hintze BJ, Lewis SM, Richardson JS, Richardson DC (2016) MolProbity's ultimate rotamer-library distributions for model validation, *Proteins: Struc Func Bioinf* **84**: 1177-1189

FAQ

Are the defaults the best for refinements?

Of course, the answer is no. The defaults have been chosen to provide the best results in the shortest time for the majority of models and data.

One of the first options to change is `optimize_xyz_weight`. Setting this option to true will optimize the weight used between the geometry and data terms of the refinement target function. Details can be found in Afonine et al., 2011.

One can also increase the number of refinement macros cycles using `number_of_macro_cycles` to ensure convergence. Ten is an adequate number.

References:

Afonine, P. V., Echols, N., Grosse-Kunstleve, R. W., Moriarty, N. W. & Adams, P. D. (2011). *Comput. Crystallogr. Newsl.* **2**, 99–103.

MARCO: The Machine Recognition of Crystallization Outcomes

Andrew E. Bruno[a], Patrick Charbonneau[b], Janet Newman[c], Edward H. Snell[d], David R. So[e], Vincent Vanhoucke[e], Christopher J. Watkins[f], Shawn Williams[g] and Julie Wilson[h]

[a]*Center for Computational Research, University at Buffalo, Buffalo, New York, USA*
[b]*Department of Chemistry and Department of Physics, Duke University, Durham, N. Carolina, USA*
[c]*Collaborative Crystallisation Centre, CSIRO, Parkville, Victoria, Australia*
[d]*Hauptman-Woodward Medical Research Institute and SUNY Buffalo, Department of Materials, Design, and Innovation, Buffalo, New York, USA*
[e]*Google Brain, Google Inc., Mountain View, California, USA*
[f]*IM&T Scientific Computing, CSIRO, Clayton South, Victoria, Australia*
[g]*Platform Technology and Sciences, GlaxoSmithKline Inc., Collegeville, Pennsylvania, USA*
[h]*Department of Mathematics, University of York, York, United Kingdom*

Correspondence email: vanhoucke@google.com

Introduction

Robust identification of the results from robotic crystallisation systems is vital to research in both industry and academia. Various research groups have approached the problem of crystal recognition using automated image analysis. As large, reliably annotated training can increase success rates, the largest training set of images previously compiled, comprising ~150,000 images, has been heavily used in that context. The resulting methods, however, often require time-consuming preprocessing stages, such as image segmentation and feature extraction. The machine-learning algorithms used for classification have thus far been specific to particular experimental setups and imaging systems.

The MARCO initiative

The macromolecular crystallization images collated by the Machine Recognition of Crystallization Outcomes (MARCO) consortium includes roughly half a million annotated images over different technical setups and imaging systems from five academic institutions and pharmaceutical companies (Figure 1. Images available from https://marco.ccr.buffalo.edu/). In contrast to the carefully curated datasets used previously, the MARCO dataset includes images with very different fields of view, problems with focus or illumination as well as those with dispensing errors. The scoring protocols of different institutions varied and, in order to homogenize the MARCO dataset, annotations were simplified to a four-class system: Crystals, Precipitate, Clear and Other. In collaboration with researchers from Google Brain, state-of-the-art deep learning algorithms were then applied to the MARCO dataset for classification. These algorithms employ Convolution Neural Networks (CNN), which require minimal preprocessing and are particularly suited to image analysis. Using a single model with all data sources combined, the trained CNN was able to correctly label 94.5% of the independent test images, regardless of their experimental origin. The algorithm and results are described in PloS one (https://arxiv.org/pdf/1803.10342.pdf) and an open source version of classifier is available at https://github.com/tensorflow/models/tree/master/research/marco.

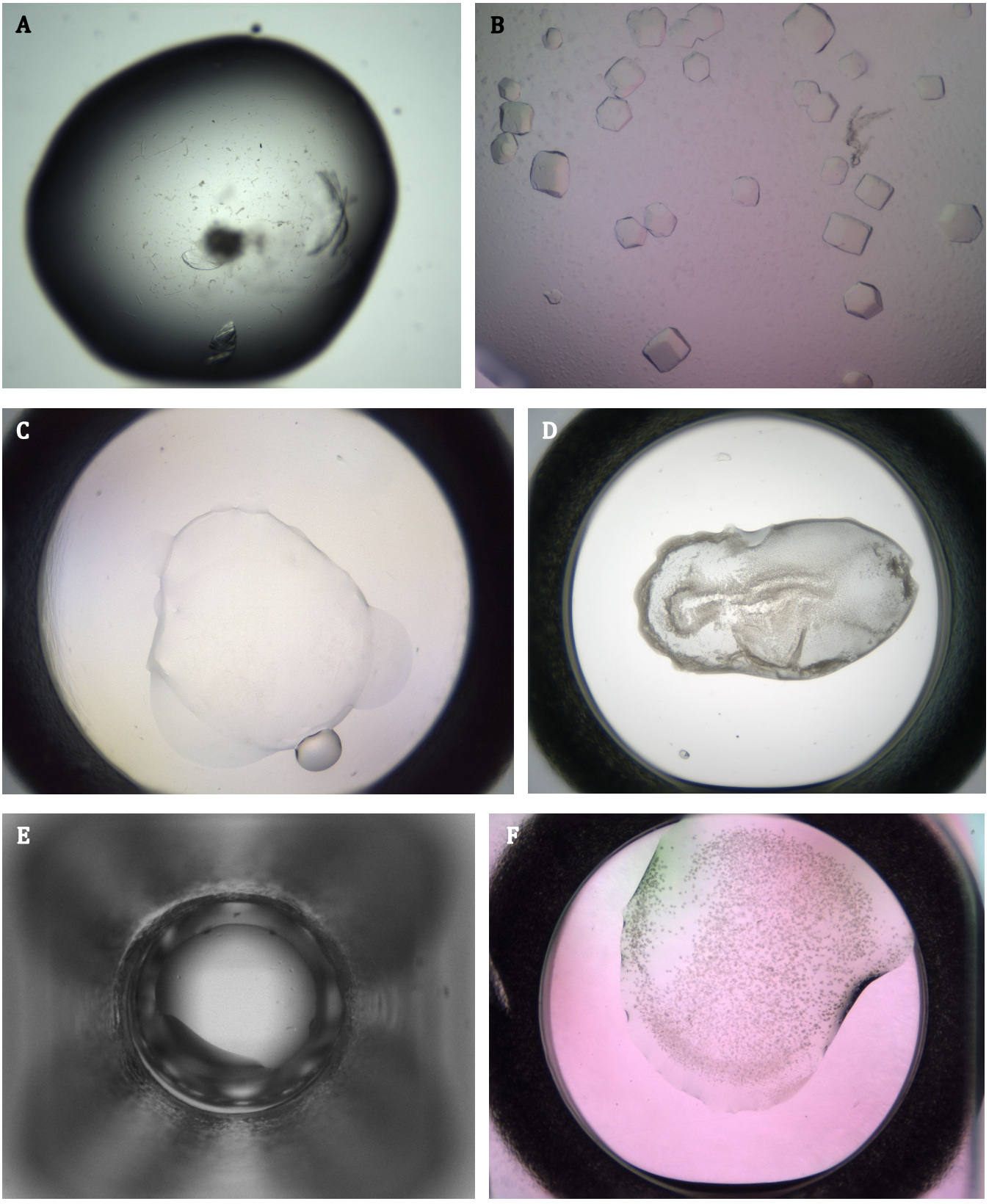


Figure 1: Images in the MARCO dataset show different experimental set-ups with various fields of view and resolutions from five industrial and academic partners: (A) Collaborative Crystallisation Centre; (B) and (F) GlaxoSmithKline; (C) Merck & Co; (D) Bristol-Myers Squibb; (E) Hauptman-Woodward Medical Research Institute.

Building a model the way you do: Map-to-model version 2

Tom Terwilliger

Los Alamos National Laboratory, Los Alamos NM 87545
New Mexico Consortium, 100 Entrada Dr, Los Alamos, NM 87544

Wouldn't it be nice if Phenix could trace the density in a cryo-EM map the way you do: find the clearest density, trace the chain at a high contour level, then dial down the contours until connections appear and trace the remainder of the chain? Seems so easy that even a computer could do it this way.

Well now it does! Version 2 of map-to-model uses a super-quick new algorithm for model-building the mimics how you would do it yourself.

Before tracing the chain, map-to-model finds helices and strands in the map. These secondary structure elements are often very accurate, so they are going to be used as fixed parts of the model to be built.

Next, map-to-model traces the chain just as you would using the new tool called trace-and-build. The trace-and-build tool chooses good density marked by the helices and strands and finds other segments of density that are very clear. Then it tries to join pairs of good segments of density by finding the highest contour level that just allows a connection. If the connection doesn't branch and isn't already used, the pair of segments is joined to make a single longer segment. This process of chain tracing is continued until no clear connections exist.

Figure 1 shows the chain tracing obtained from the small rotavirus map provided in the Phenix distribution as a model-building example. The map-to-model algorithm finds one long chain and a few short fragments. When you run map-to-model in the Phenix GUI, this chain tracing with the density and path of each chain displayed for you automatically.

Once the path of a chain is identified, a protein model is built using that path as a guide. Imagine you are building a model and you have traced the chain. What is the next thing you are going to look for? Surely you'll look for side chains marking the C_β positions. Once again map-to-model does this just the way you would. It looks for density coming off the path of the main-chain and marks likely C_β positions. Then it uses the new tool called `refine_ca_model` to find a set of C_α and C_β positions that are spaced 3.8Å apart and that match the likely C_β positions as closely as possible. At this point the helices and strands that were identified at the beginning of the procedure are spliced into the chains, creating a mosaic model and using the fixed secondary structure elements wherever they are present. With the C_α and C_β positions for a chain identified, map-to-model uses the tool Pulchra (Rotkiewicz & Skolnick, 2008) to generate an all-atom model that is refined against the map with the Phenix real-space-refine tool.

The last step in model-building is to figure out what part of the sequence in your sequence file is associated with each segment in the model. This is done in map-to-model with the new tool called `sequence_from_map`. At each position in the model, the density in the map

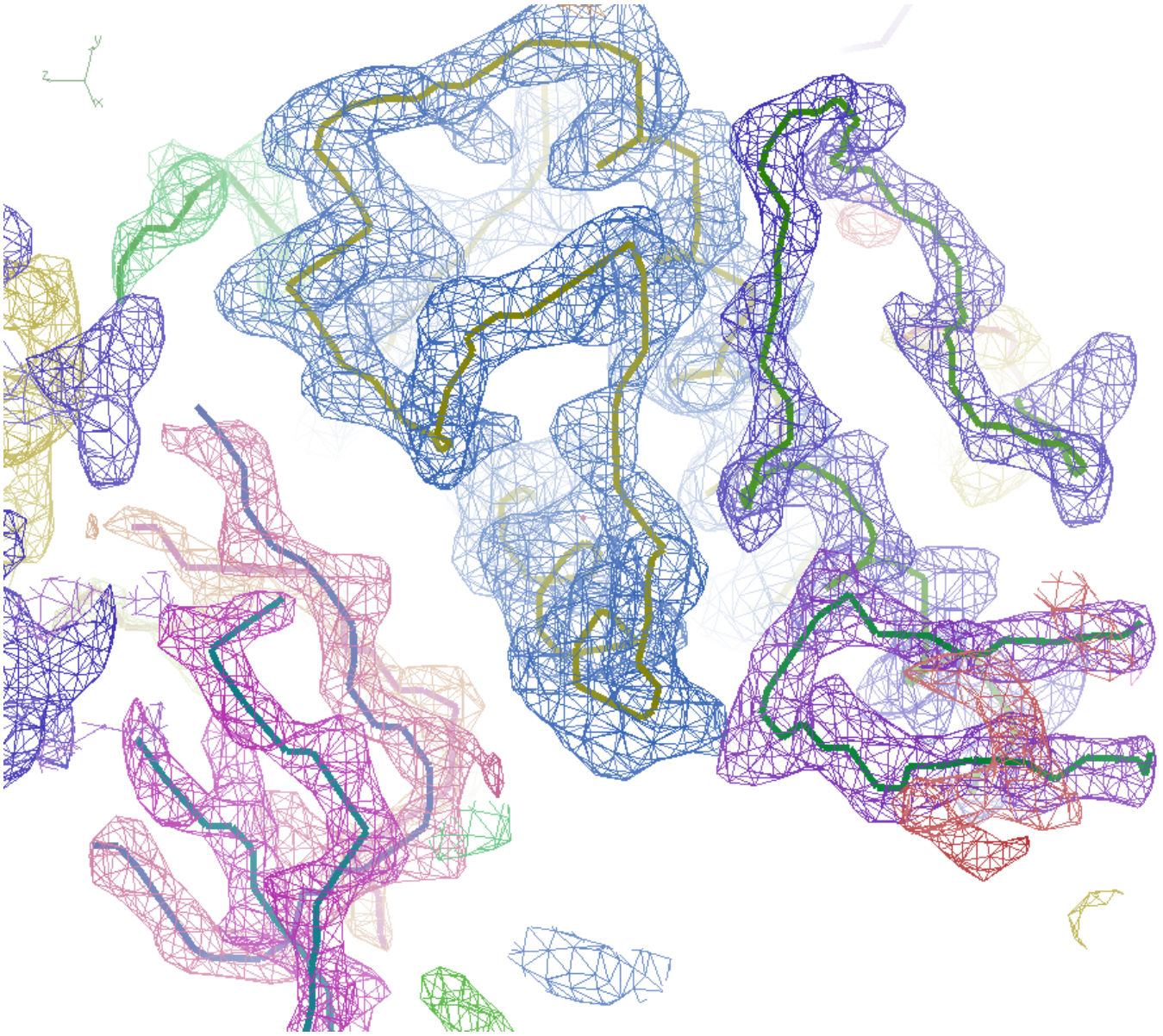


Figure 1: Example of chain tracing.

at the side-chain position is compared with expected density for each rotamer of each possible amino acid, and a relative probability for each amino acid at each position is calculated. A pseudo-sequence is then created using the most likely amino acids at each position in the model. This map-based sequence is then aligned to the supplied sequence and the best alignment is chosen and the corresponding amino acids are used at each position in the model.

Figure 2 shows part of the model created in this way for the small rotavirus map used in figure 1. The entire process takes about 5 minutes on a 4-processor machine for this small structure. The model is not perfect (it has some insertions/deletions) but it is very close to the known structure of this rotavirus protein.

Once you have built a quick model with map-to-model, you can go back and improve it. If you can see that some segments really should

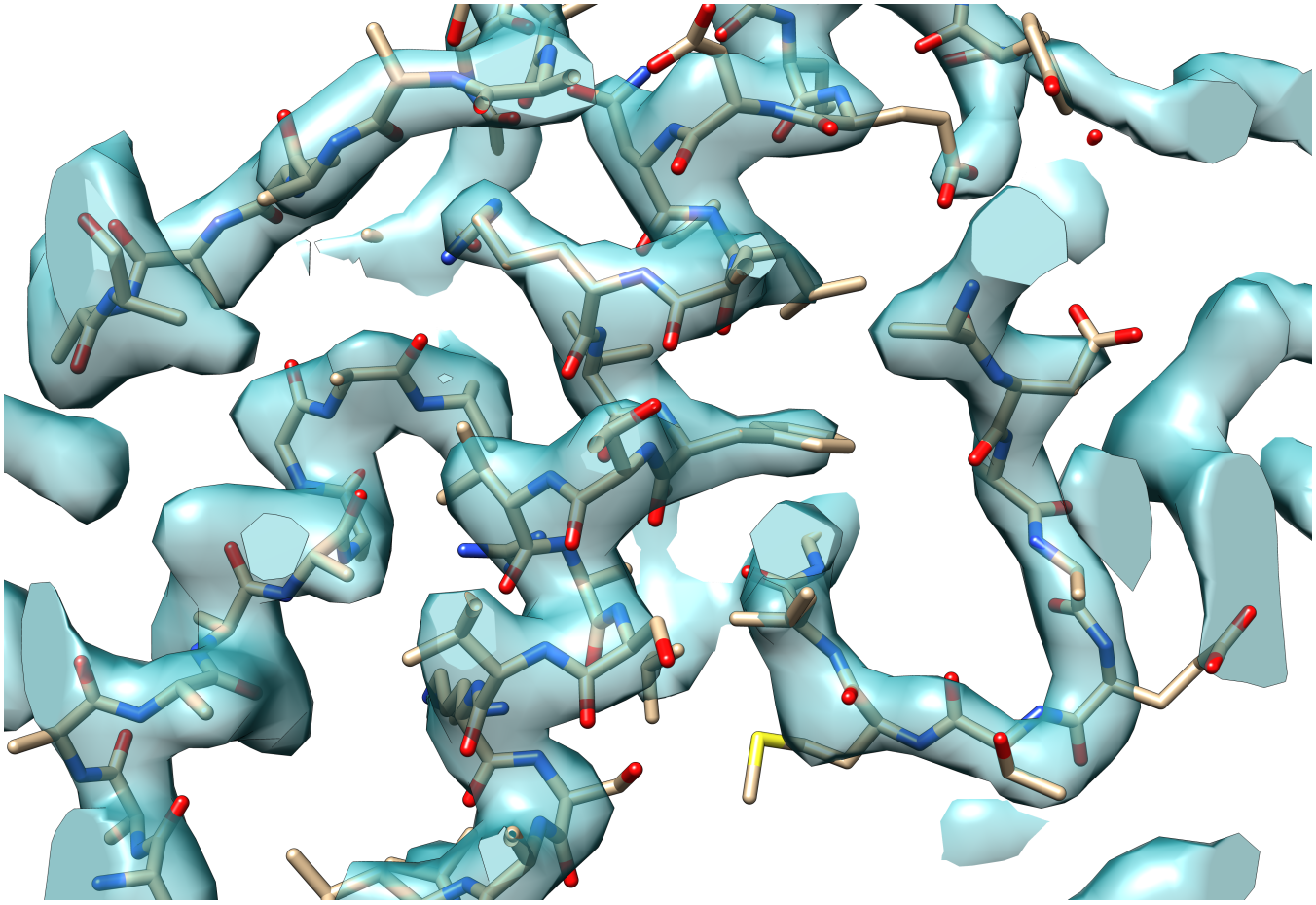


Figure 2: Model created from tracing in figure 1.

be joined, you can feed just those segments back into map-to-model and tell it to connect them. Or if you want to get rid of some of the sequence errors, you can feed your model back into map-to-model and tell it to run fix-insertions-deletions. It will use the sequence to try and identify where insertions and deletions are present and it will rebuild those segments with the appropriate number of residues.

Give the new map-to-model a try and let us know of anything that you would like improved!

Reference:

Rotkiewicz, P., J. Skolnick. J. (2008). Fast procedure for reconstruction of full-atom protein models from reduced representations. Comput Chem 29, 1460-5.