Gradient-flow Segmentation of Crystallographic Maps

A. Fyfe[a]* and W.G. Scott[a]

[a] *University of California, Santa Cruz, USA. E-mail: afyfe@soe.ucsc.edu*

Abstract

Continuous scalar functions defined over a three-dimensional domain can, under fairly general conditions, be used to segment the domain into regions of uniform gradient flow. This mathematical device has found widespread application in various fields including data visualization, computational chemistry, topography and image analysis. For such applications, diverse algorithms applicable to grid-sampled data are available. Here we extend the technique to accommodate domains subject to the constraints of space group symmetry and apply the result to the interpretation of crystallographic density maps. Standard tools for examining the features of crystallographic maps include peak/pit searches, inspection of isovalue contours, skeletonization and medial axis extraction. We demonstrate that segmentation of maps into disjoint basins characterized by uniform gradient flow provides a valuable addition to the crystallographer's toolbox. A distinguishing strength of this method is the ability to identify and inspect the map volume over which a modeled atom contributes the dominant fraction of total density.

1. Introduction

Crystallographic model building is an iterative process: a hypothetical model is formulated, the electron density for the model and the structure factors that correspond to the density's representation in reciprocal space are calculated, and the agreement between the amplitudes of the hypothesized and observed structure factors measured. Until the model gives an acceptably good fit,as determined by R-factor statistics, the model is adjusted and the process is repeated. Though R-factors calculated in reciprocal space directly assess global agreement between a proposed model and the experimental data, they are notoriously unhelpful in diagnosing local discrepancies. Thus, a good deal of crystallographic practice is taken up with the examination of maps, scalar data sets defined over $\mathbb{R}^3$ that encode various functions of the electron density, ρ, of a proposed model.

Even when restricted to their smallest unique subset, the asymmetric unit (ASU), maps are large and noisy data sets so various tools have been developed to help crystallographers extract salient features. Tools in widespread use include lists of high and low points, isovalue contours, and map skeletons (Zhang *et al.*, 2001). Here we investigate the applicability of tools based on segmentation of maps into disjoint basins defined by uniform gradient flow. In such segmentation, each basin gathers all points in $\mathbb{R}^3$ for which the density's gradient vector, $\nabla\rho$, points to the same local maximum. Figure 1 illustrates segmentation of a two-dimensional domain on the basis of the height function.

The algorithms and terminology underlying decomposition of a scalar function on $\mathbb{R}^3$ into regions of uniform gradient flow rely on concepts that originate in a setting where ρ is a well-behaved, continuous and differentiable function. In application, these concepts are transferred to equivalent approximations on the grid-sampled data available for

analysis. The gradient of the scalar field sampled by the map, $\nabla\rho$, defines a vector field over $\mathbb{R}^3$. A curve $l(s), \mathbb{R} \to \mathbb{R}^3$ defined such that the tangent at each point in the curve coincides with the gradient vector at the point,

$$\frac{\partial}{\partial s} l(s) = \nabla\rho(l(s)) \tag{1}$$

is an *integral line* (Zomorodian, 2009). Gradient vectors are null at critical points of ρ, thus an integral line describes a path that originates and terminates at a critical point of ρ. From their definition, two integral lines are disjoint or identical, except possibly at their endpoints, and the union of points in all integral lines covers the domain of ρ. It follows that sets of integral lines that share a common origin or terminus may be defined, and that such sets may be used to partition $\mathbb{R}^3$. The set of points associated with integral lines that terminate at a maximum critical point $\boldsymbol{p}$ define the *descending* or *stable manifold* of p. Similarly integral lines that originate from a minimum define the *unstable* or *ascending manifold* of $\boldsymbol{p}$ (Zomorodian, 2009). In a general setting, construction of disjoint gradient-flow regions depends on additional technical requirements: the function's critical points must not be degenerate and the stable and unstable manifolds must intersect transversely (Matsumoto, 1997). On actual data these requirements are easily met, by minor perturbation if necessary. For a height function as shown in Figure 1 for example, the surfaces defined by the ascending and descending basins of distinct critical points are required to meet at curves with distinct points of intersection.

The ideas and methods underlying volumetric segmentation of a domain on the basis of uniform gradient flow are well established. In mathematics, *Morse Theory* exposes relationships between the critical points of a function and the topology of the domain it is defined on (Matsumoto, 1997). In theoretical chemistry, segmentation has been used to recover the boundaries of constituent atoms within a density distribution

(Bader, 1990). In data visualization, construction of Morse-Smale complexes has been exploited for smoothing and feature extraction in large, complex data sets (Gyulassy *et al.*, 2011). However the method has not yet found widespread use in crystallographic practice.

A variety of algorithms for the computation Morse-Smale complexes on grid-sampled data have been devised; they can be loosely grouped into three classes. The mathematical results of Morse Theory, though initially defined for smooth functions, are also applicable to polyhedra,as shown by Banchoff (Banchoff, 1970). In the data visualization community, influential early work by Edelsbrunner and colleagues provided algorithms for extending constructs originally defined for smooth functions to piece-wise linear manifolds (Edelsbrunner *et al.*, 2003). Initially developed for the two dimensional case, these algorithms were later extended to three dimensions by Gyulassy and others (Gyulassy *et al.*, 2007). The approach we use here relies a variant of this method. A second class of algorithms has arisen from a formulation of Morse Theory in terms of cell complexes, intrinsically discrete objects with no reference to the continuous case. The discrete formulation of Morse Theory, introduced by Robin Forman (Forman, 1998), has been followed by algorithms that reinterpret segmentation of grid-sampled data in terms of operations on cell complexes. Henry King and coworkers first developed an algorithm for constructing discrete Morse complexes for grid-sampled data (King *et al.*, 2005) and this has been followed by others (Gyulassy *et al.*, 2008), (Robins *et al.*, 2011). A third class of algorithms originated in the theoretical chemistry community as numerical solutions to the construction of *zero flux surfaces* defined by Bader's theory of *Atoms in Molecules* (Bader, 1990). Though these algorithms have evolved with no direct reference to discrete or continuous Morse Theory, their intent is the construction of boundaries within a group of atoms defined by descending basins of electron density. A popular algorithm by Henkelman and cowork-

ers used for charge analysis effectively traces integral lines to define a descending basin (Henkelman *et al.*, 2005),(Tang *et al.*, 2009). Malcolm and Popelier showed that the widely used *octree* algorithm of computer graphics can be successfully adapted for delineating basin boundaries (Malcolm & Popelier, 2003).

Among the above algorithms there are consequential differences and trade offs. For example, formulating the problem as an operation on cell complexes yields significant improvements in performance and readily extends to multi-dimensional domains but discards important geometric information. Here we are primarily interested in exploring applicability of gradient flow segmentation to the interpretation of crystallographic maps. Restricting the domain of interest to the crystallographic asymmetric unit reduces the size of the problem to one readily handled by any of the above algorithms. Nevertheless, novel adaptations are needed to handle symmetry restrictions of the domain being partitioned and the idiosyncrasies of map data sets.

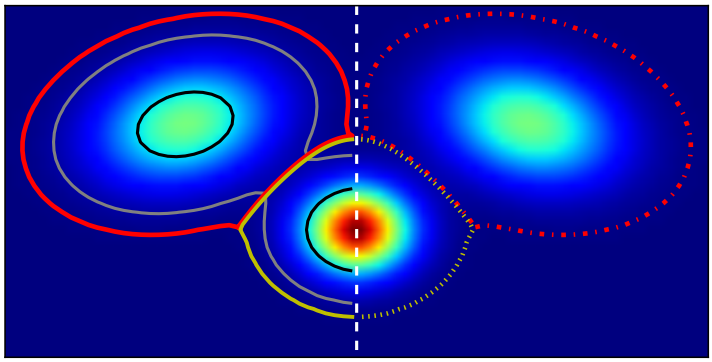


Fig. 1. *Gradient-flow segmentation of simulated density for a unit cell in plane group pm. The magnitude of simulated density from two point scatterers partitions regions within the two asymmetric units of the cell. Contours lines are shown as thin lines in black and gray. Descending basins for each of the two point scatterers are shown as thick lines in red and yellow. Dotted lines show the basin boundaries outside the ASU. See text for interpretation.*

2. Methods

A schematic overview of our map segmentation algorithm is shown in Figure 1. The two asymmetric units of a unit cell in plane group *pm* are filled with simulated density from two point scatterers. The white, dashed, line denotes the symmetry element. Contour lines at two levels are shown as thin lines. The higher, black, contour clearly distinguishes the two scattering sites but omits much of the area at lower density. A contour with lower isovalue, shown by the thin gray line, captures much of this area but does not distinguish the two sites. Gradient-flow segmentation yields two descending basins, shown as thick lines in red and yellow. Each basin captures the

largest area of the cell over which the dominant component of the total density is contributed by density descending monotonically from the basin maximum. Of the two basins, the one shown in red is interior to the ASU whereas the one shown in yellow is split by the ASU boundary. For analysis, the complete extent of this basin is reconstructed via the symmetry relation. Only the ASU is processed by the algorithm, the basin boundaries for the remainder of the unit cell are shown as dotted lines.

As applied to an actual crystallographic map, the algorithm entails four stages of processing. Starting with an encoding of the ASU in standard CCP4 (Winn *et al.*, 2011) map format, we first obtain a grid suitable for segmentation. In this preprocessing step, the data can optionally be re-sampled by a user-specified factor. This expansion accommodates the later partition of map voxel points into basin and boundary points. We also sort voxel values so as to ensure a strict order; ties in the original density values are broken by reference to map indices (Robins *et al.*, 2011).

The second stage constructs the three-dimensional basins and their two-dimensional separating boundaries. Here we adapt a method introduced by Gyulassy (Gyulassy *et al.*, 2007), though a different approach is used to compute the two-dimensional surfaces separating basins. For each grid point $\mathbf{p}$, a key data structure in the computation is $Link(\mathbf{p})$, the set of neighboring points and the related sets $Link^{-}(\mathbf{p})$ and $Link^{+}(\mathbf{p})$, the subsets of $Link(\mathbf{p})$ whose values are smaller or larger than $\mathbf{p}$. This cluster approximates a spherical neighborhood of $\mathbf{p}$ and is implemented as the 27-point sub-grid centered at $\mathbf{p}$. A subset S of $Link(\mathbf{p})$ is considered connected if for each point $s \in S$ at least one grid point in the six-point neighborhood of s is also in S. Intuitively, the neighborhood of a point interior to a basin will contain only other interior points or boundary points, whereas the neighborhood of a boundary point will include points interior to at least two basins. Visiting grid points in sorted order ensures a point is

classified as interior if all its neighbors flagged as interior have higher density values, in a manner similar to the watershed algorithm (Roerdink & Meijster, 2000).

The set of local maxima of the map is found as the set of voxels for which $Link^{-}(\mathbf{p})$ is empty; these define the seed points for each basin. Following Gyulassy (Gyulassy *et al.*, 2007). data values are examined in sorted order and tested for membership as interior to a basin. If $Link^{+}(\mathbf{p})$ is connected and all elements of $Link^{+}(\mathbf{p})$ are classified as interior to a given basin, then $\mathbf{p}$ is also considered interior to that basin.

Once the collection of basins and their interior points is computed, a second pass screens remaining points. Points not labeled interior but whose link includes interior points are classified as boundary points. For each boundary point, we track the set of basins it is incident on and, for each basin, the set of boundary points incident to it. The list of unique basin incidences, across all boundary points, serves to label each surface instance. That is, each surface separates a unique set of basins. For each surface, membership is obtained by computing the set intersection of the incident boundary points of all basins defining the surface. On conclusion of this step, all points in the original map are partitioned into disjoint subsets consisting of basins and their separating surfaces.

The third stage of processing amends the derived partition to incorporate the symmetry of the space group assigned to the map. Basins are classified as interior or boundary, depending on whether the link of any interior point is incident on a face of the asymmetric unit. Among boundary basins, we identify those whose critical point, $\mathbf{cp}$, is also incident on a face and inspect the basin membership of $Link(\mathbf{cp})$. Let $\mathbf{BB}$ be such a boundary basin. Since $Link(\mathbf{cp_{BB}})$ includes points outside the asymmetric unit, the basin membership of such points is determined by applying the space group's

symmetry transformations to obtain the equivalent point within the asymmetric unit. Let n be the number of points in $Link(\mathbf{cp_{BB}})$ interior to basins other than **BB**. If **n** is zero, no adjustment is needed since $Link(\mathbf{cp})$ contains only points interior to **BB** or boundary points. If **n** is one, **BB** is merged to join the corresponding basin. In this case, gradient flow defines a single basin whose separation into two parts is introduced by the conventional choice of asymmetric unit. If **n** is greater than one, we reclassify **cp** as a boundary point, find the new local maximum of **BB** and repeat the previous analysis.

Having computed a symmetry-adjusted segmentation we are interested in the properties of the resulting basins. These include the basin critical point, the basin volume, the total density integrated over the basin and the correlation between computed and observed density over the volume defined by the basin. The final stage of processing constructs a data structure that supports such queries and enables quick mapping of Cartesian, grid, or fractional coordinates to the corresponding basin or boundary.

3. Results

We apply the techniques introduced in the previous section to different types of maps commonly computed in the course of crystallographic analysis. Maps can be grouped into two classes, *density maps* expected to be everywhere non-negative, except for scaling adjustments, and *difference maps* expected to take both positive and negative values near zero. Segmentation of the two classes is approached somewhat differently. For density maps, we ensure non-negative values by subtracting the map minimum then partition the map into ascending and descending basins. Optionally, basins can be further aggregated, for example by combining all descending basins with peaks below a specified level and outside the molecular surface into bulk solvent basins. For

difference maps, segmentation is also applied twice, but in a different manner : first by computing ascending basins over positive map values, then descending basins over negative values. This approach allows us to separately identify basins of positive and negative density values.

As concrete examples, we focus on two structures deposited in the PDB databank (Berman *et al.*, 2003), 1OK0 and 3ZP8. The former is the structure of a small, single chain, 74 residue, protein to 0.93$\mathring{A}$ resolution refined to R_{work}/R_{free} values of 10.3/13.0 with *shelxl* (Sheldrick, 2008) by A. König and co-workers (König *et al.*, 2003). It is a particularly apt example for our purpose since its refinement was chosen by T. Schneider as a worked example of the application of *shelxl* to macromolecular refinement; thus all files for reproducing and inspecting successive refinement stages in the hands of expert practitioners are available (Muller, 2006). There are minor differences between the deposited version of 1OK0 and the final iteration of this refinement series, "tenda12_cgls"; throughout, our analysis refers to the latter.

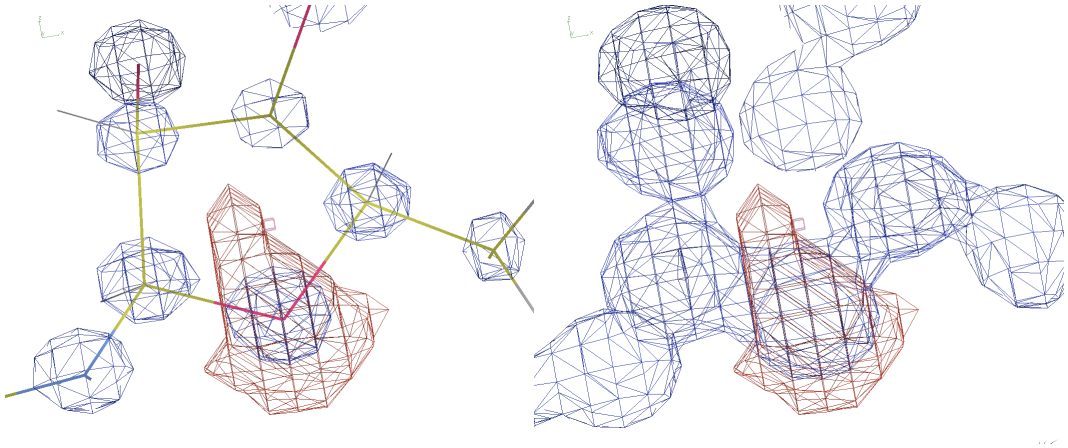
The second example is the structure of a full-length RNA hammerhead ribozyme to a resolution of 1.55$\mathring{A}$, recently deposited by our laboratory (Anderson *et al.*, 2013) The structure was refined to R vales of 18.9/21.6 with *phenix.refine* (Adams *et al.*, 2010). The two chains of the structure, substrate and enzyme, are composed of 43 and 20 residues respectively. To inactivate the enzyme, cytodine 6 of the substrate strand, labeled C17 in canonical hammerhead active site numbering, includes a methyl group on the ribose O2'.

In both examples, we focus on map segmentation as a diagnostic tool in the placement of ordered solvent. Distinguishing electron density attributable to solvent from background noise remains challenging and the stereochemical "prior knowledge", such

as permissible torsion angles, bond lengths, chiral volumes and planarity constraints, that guides the interpretation of electron-density maps and their transformation into reliable atomic models, is not available for ordered solvent. The final refinement of structure 1OK0 includes 159 waters, including some placed at alternate positions with partial occupancy, and two Cl^- ions, all modeled anisotropically. Structure 3ZP8 includes 289 waters and 16 Na^+ ions, all at full occupancy unless incident on special positions and modeled isotropically.

3.1. Basin Properties

Ideally, segmentation of the density calculated for a model would yield the boundaries of each atom in the hypothesized molecular arrangement. In practice, as discussed in the next section, thermal displacement, resolution and disorder attributable to unmodeled effects, limit the granularity of segmentation so that multiple covalently bonded atoms often share the same basin. Where atomic coordinates coincide with a local peak , a unique basin will be defined as density values descend to the corresponding zero flux surface.



(a) (b)

Fig. 2. *Descending basin of a ribose O4' atom. The descending basin calculated for a ribose O4' atom from segmentation of a calculated, DFc, map is shown in red alongside isovalue surfaces contoured at (a)* 3σ *and (b)* 1.5σ

Figure 2 contrasts the descending basin for a ribose O4' atom from a calculated density map in red along with isocontour surfaces in blue drawn at 3σ of the map rmsd in panel(a) and 1.5σ in panel(b). Inspection reveals some key differences between basin surfaces and isovalue contours:

- Unlike isovalue contours, basin surfaces enclose a single peak, regardless of the map variance. This property also applies to isovalue contours taken at sufficiently high values of σ but does not hold at lower levels as shown in panel (b). The contours of panel(b), more representative of noisier, lower-resolution, maps, include multiple atom peaks and thus the enclosed volume is less useful for calculating local, atom-specific, properties. As further discussed in the next section, the granularity of basins relative to the atomic model is an intrinsic property of local features of the map used for segmentation whereas, for isovalue contours, granularity is a function of the map's global variance.
- Whereas all points on the surface of an isovalue contour, by definition, have the same value of ρ, density values on the surface of a basin vary widely. The

highest values occur near points on the vector joining covalently-bonded atoms and lower values where the basin adjoins unmodeled bulk solvent.

- Typically, the map volume enclosed by an atom's basin is considerable larger than the volume enclosed by the surface of the smallest isovalue contour uniquely associated with the atom, a property that follows from the the defining gradient-flow characteristic of basins. As one travels outward from an atom's density peak, values of ρ decrease. As this path approaches a neighboring modeled atom, an abrupt boundary will separate the two basins. However, if the path points towards bulk solvent, the distance to the next local maximum will be greater and all points along this path will be assigned to the original basin. Thus basins generally expand towards unmodeled regions of lower density.

This effect is apparent in Figure **??** in the sharp boundaries that separate the O4' atom from the adjacent C3' and C5' atoms. Figure 3 illustrates the effect more globally as it applies to non-bonding interactions among 284 of 289 water molecules in structure 3ZP8 assigned to single basins by segmentation of the final DFc map. The basin volume, in grid voxels, is plotted against the number of neighboring atoms within plausible distance for non-bonding interaction, 2.3$\mathring{A}$ to 3.5$\mathring{A}$, including symmetry equivalents. Points are colored according to the real space correlation calculated over the basin between the DFc and 2mFo-DFc maps. The inverse relationship between volume and the number of neighbors illustrates how basins expand into volume modeled as bulk solvent. Furthermore, larger basins with fewer neighbors, generally exhibit poorer agreement between calculated and observed density, possibly highlighting a deficiency in the modeling of density for regions further from the atomic surface. The variance among basin volumes is inversely related to the number of neighboring atoms.

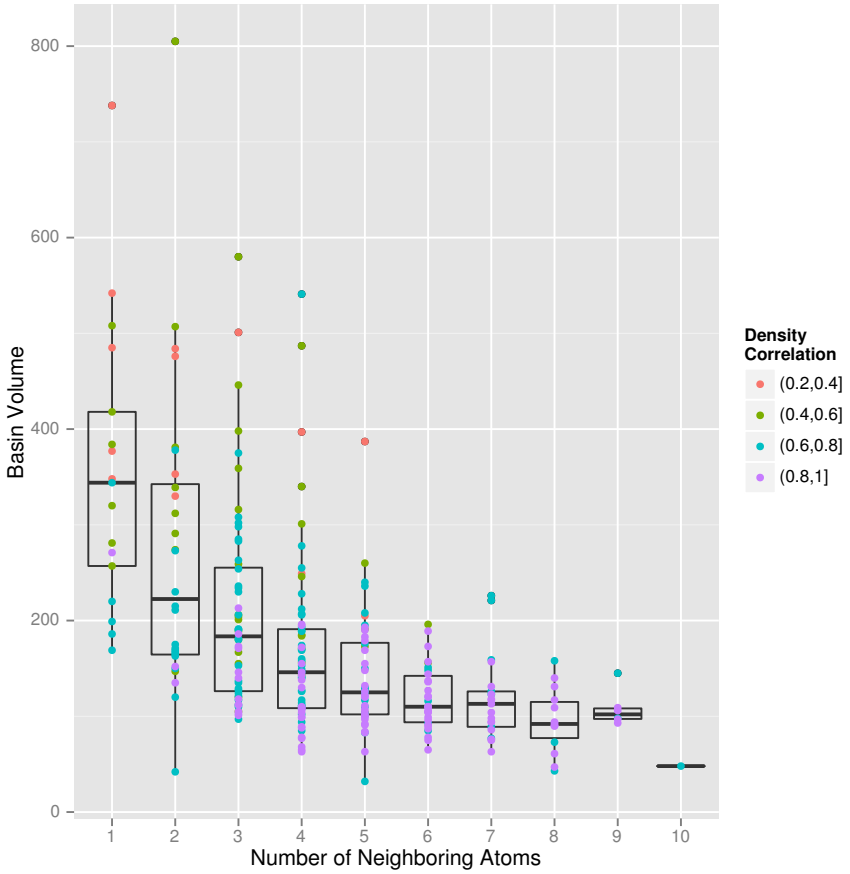


Fig. 3. *Basin volume versus crowding among ordered solvent. The volume of descending basins, in grid voxels, of modeled water molecules for structure 3ZP8 is plotted against the number of possible non-bonding contacts. The box plots for basin volume show bars for the first, second and third quartiles. Points are colored by value of the correlation between calculated and observed density, as detailed in the text.*

All maps for a structure have the same crystallographic properties and generally the same grid spacing; they differ only in the scalar quantity estimated. We typically rely on volumetric segmentation of the computed, DFc, map to obtain basins for the hypothesized model and then use the grid region circumscribed by each basin to examine the local relationship between scalar quantities estimated by different maps. A variety of indicators computable from this approach can provide diagnostic information regarding the maps:

- Basin Density Distribution. For each basin we can compute a volume and for each type of map we can obtain standard measures of the distribution of the scalar quantity involved such as the basin peak, variance, skewness and kurtosis. Since calculated density for different elements decreases at different rates, spherically-symmetric elevation centered near an atomic coordinate in the residual, $mFo - Dfc$, map can cast doubt on the identity of the atom modeled for a solvent basin. Integrating density over the basin provides information about "crowdedness" of the atom's neighborhood since, as noted, basins of atoms expand into bulk solvent in the absence of bonding interactions and thus exhibit larger volume and greater total density.
- Correlation. The Pearson product-moment correlation ρ_{F_C,F_O} between observed and calculated density estimates is a key measure of the agreement between the data and hypothesized model. Good agreement, as expressed by ρ_{F_C,F_O} values greater than 0.8, indicates confidence in the type of scatterer and B factor modeled for the basin. Agreement between calculated and observed maps supports the hypothesized model and the extent of agreement can be separately assessed for its constituent parts (Rupp, 2006). One would expect R-factor statistics, as global indicators of model quality, to agree with distributions of atom by atom correlation between calculated and observed density. This is confirmed over twelve successive refinements of structure 1OK0 as shown in panel (a) of Figure 5. However, where the ρ_{F_C,F_O} value is low, this statistic does not provide information about the source of the discrepancy.
- Spatial Auto-correlation. If the hypothesized model fits the data, the residual density map with coefficients $mFo - Dfc$ should consist of featureless error, or "white noise", independent of location. A statistically unlikely clustering of higher or lower values within a basin suggests a shortcoming of the model. Fourier

truncation ripples yield spurious local maxima, thus segmentation that produces small basins spherically distributed about a modeled heavy atom suggests a problem in density estimation. Independence between magnitude and location in residual maps can be tested by spatial auto-correlation estimators such as Moran's I or Geary's C, which, though widely used in image processing and geographic information systems (Marwan *et al.*, 2009) are less commonly applied to crystallographic maps.

3.2. Basin Granularity

The disordering impact of thermal displacement factors, partial occupancy and incomplete modeling compound the loss of detail intrinsic to lower resolution and dampen peaks of calculated density estimates obtained from standard Gaussian approximation. The effects are well characterized and can be summarized graphically (Grosse-Kunstleve & Bourhis, 2011; Afonine & Urzhumtsev, 2004); their impact limits instances where the basins of a segmented map are associated with an individual atom.

As a specific example, we consider density in the neighborhood of the P-O1P segment of a phosphodiester bond. Structure 3ZP8 includes 61 P-O1P pairs with a mean distance of 1.487$\mathring{A}$ and standard deviation of 0.005. The average isotropic B factors are 35.686$\mathring{A}^2$ (s.d. 11.486) and 39.987$\mathring{A}^2$ (s.d. 12.665) for P and O1P respectively. Figure 4 plots the calculated density for bonded P and O atoms with those parameter values obtained from a 5 term Gaussian approximation with coefficients from the International Tables for Crystallography, Volume C (1992) (IUCr, 1992). Density as a function of radial distance is calculated as (Grosse-Kunstleve & Bourhis, 2011; Afonine & Urzhumtsev, 2004):

$$\rho(r) = \sum_{i=1}^{N} a_i \left(\frac{4\pi}{b_i + B_{iso}} \right)^{3/2} \exp\left(-\frac{4\pi^2 r^2}{b_i + B_{iso}} \right) \qquad (2)$$

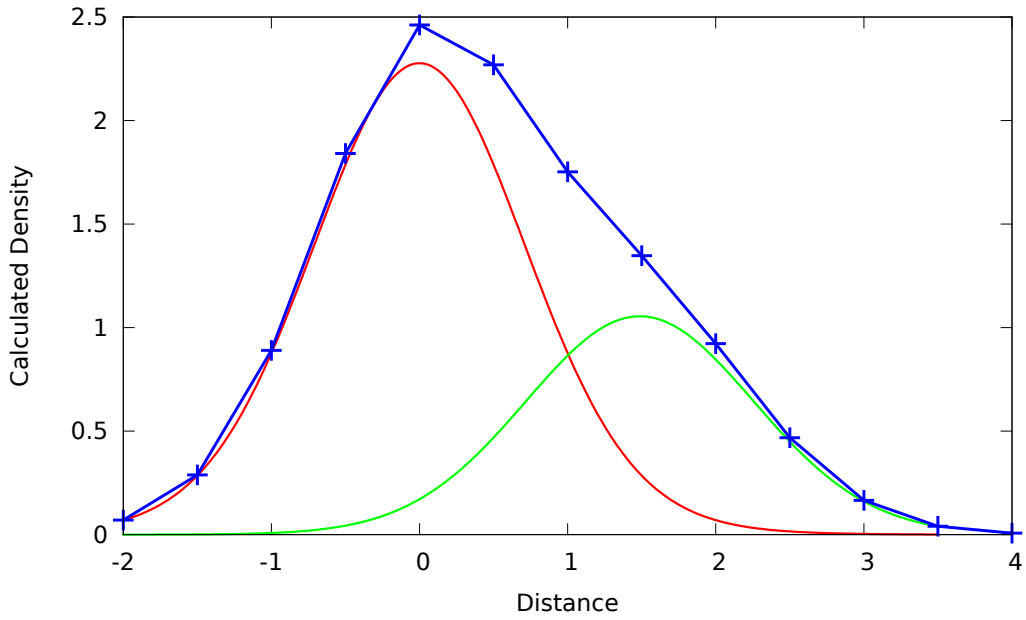


Fig. 4. *Calculated density along a P-O1P bond. Density for the P, in red and centered at the origin, and O1P, in green and displaced along the x axis, atoms of a phosphodiester bond was calculated as detailed in the text. Also shown, in blue, is the density from a FFT-based structure factor calculation to 1.55Å. Density is calculated in* $e/Å^3$ *and distance along the bond is in Å. Blue crosses correspond to map grid points.*

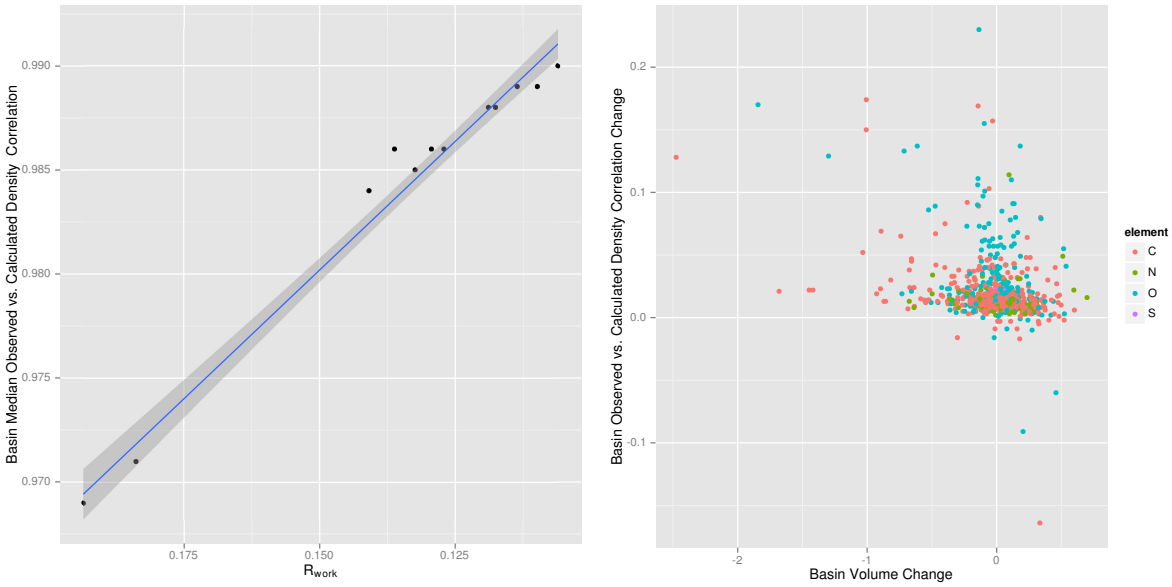
Along with the "exact" atom density, the plot shows the density recorded in a map obtained from Fourier synthesis of the calculated structure factors from a simple model representing the two atoms. In a cube of side $10Å$ in space group P1, the P atom was placed at position (5,5,5) and the O atom was displaced by $1.487Å$ along the x axis. Structure factors for this model were calculated to a resolution of $1.55Å$ with the standard FFT-based algorithm as implemented in the *cctbx* libraries (Grosse-Kunstleve *et al.*, 2006) and the *sfall* program (Collaborative Computational Project, Number 4. 1994, 1994) and the corresponding map was calculated with the *fft* program. Though the FFT-based density approximation using the given B factors, bond distance and resolution yields values in good agreement to the exact density, the map does not provide distinct maxima for the P and O atoms.

Whether segmentation of map at a given resolution is able to resolve specific atoms depends on the topography of the electron density in an atom's neighborhood. For example, in the segmentation of the 3ZP8 structure using the calculated 1.55$\mathring{A}$ map, the phosphate O1P and O2P oxygens never resolve into individual basins, whereas 13 of the 61 modeled ribose O4' oxygen atoms occupy individual basins. The longer distances of non-bonding contacts in ordered solvent yield fewer multi-atom basins: 284 of 289 modeled waters occupy individual basins as do all 16 modeled Na atoms. Overall, of the 1649 modeled atoms in structure 3ZP8, 778 resolve to individual basins. For structure 1OK0, which benefits from higher 0.93$\mathring{A}$ resolution, among 743 non-hydrogen atoms, 625 resolve to individual basins. Among atoms that do not, 101 (87%), are involved in alternate conformations. Assignment of multiple atoms to a basin confirms that no local maxima exist at the individual atomic coordinates and thus highlights the limitations of the data in specific regions of the model.

Occasionally, atomic coordinates coincide with grid positions labeled as basin boundaries by our segmentation algorithm. This was observed in the segmentations of maps from both 3ZP8 and 1OK0 structures. The situation clearly flags a poorly resolved atom, and can be disambiguated by segmentation of a higher resolution map. In general, it should be possible circumscribe an atom's basin to a unique region by using a calculated map of sufficiently high, albeit fictitious, resolution for segmentation. In the analyses discussed here however, we have relied on segmentation of maps calculated to the resolution supported by the data.

3.3. Detecting Outliers

The principal benefit of segmenting the unit cell into regions of uniform gradient flow is the assessment of model fit on a local basis, ideally atom by atom. Overall, basin local properties can be expected to agree with global measures of quality.



(a) (b)

Fig. 5. *Global map indicators and local basin properties in structure 1OK0. (a) The reduction in R_{work} over twelve successively improved refinements of structure 1OK0 is matched by an equivalent improvement in the correlation between observed and calculated density over basins associated with modeled atoms. (b) Changes in basin volume and basin correlation, for observed versus calculated density, after the introduction of anisotropic displacement parameters in refinement.*

For example, panel (a) of Figure 5 illustrates how median basin real-space correlation tracks the R_{work} reduction obtained over twelve successively improved refinements of Tendamistat at $0.93\mathring{A}$. Inspection of local properties can reveal additional, more informative, detail beyond global fit however. The scatter plot in panel(b) shows the shift in basin correlation and volume after the introduction of anisotropic displacement parameters in the third stage of refinement, an improvement that, as evident in panel (a), yielded the largest R_{work} reduction. For each basin, volume change as a fraction of the new volume, $(V_{new} - V_{old})/V_{new}$ is plotted against the change in the density correlation over the basin.

It is evident that the two changes are not evenly distributed around zero. In accord with the sizable drop in R factors, correlation improves for 610 out of 629 basins. The change in basin volume is more symmetric, with 308 basins showing a decrease. However a more striking and unexpected pattern is that the lower left quadrant is nearly empty: a decrease in volume is rarely associated with a poorer density correlation. As basins encompass density further from the atomic peak, the agreement between modeled and observed density deteriorates. Plotting basin properties also makes it easier to detect outliers. For example the poor correlation of the basin shown at the lower right corner, corresponding to atom CD2 of leucine residue 74, is related to improperly modeled waters, corrected in later stages of refinement.

This relationship between basin volume, integrated density and correlation is also apparent in Figure 6. Here we consider the 284 single-atom basins of waters modeled for structure 3ZP8. For each basin, volume is plotted against total density from the final 2mFo-DFc map. Points are colored according to the correlation between observed and calculated density over the basin. Integrated density increases linearly with basin volume as expected; it is also apparent that more compact basins yield better agreement between observed and calculated density.

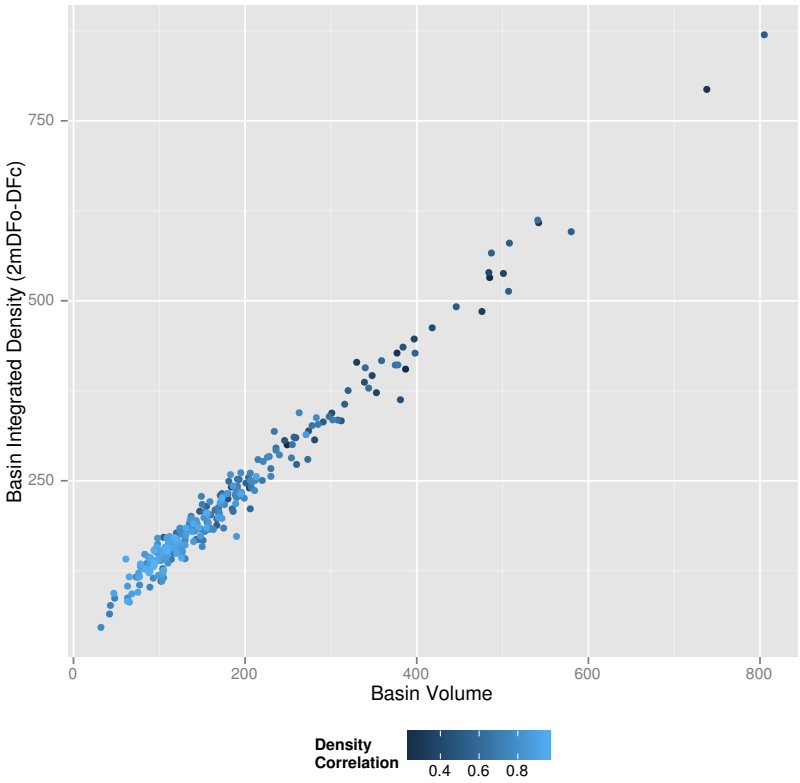


Fig. 6. *Basin volume, total density and correlation among waters for structure 3ZP8. Basin volume, in voxels, is plotted against scaled integrated density for 284 single-atom water basins. Color indicates the correlation, ρ_{F_C,F_O}, between observed, 2mDFo-DFc, versus calculated, DFc, density over each basin.*

Modeling of ordered solvent is largely driven by peaks found in the data rather by chemical expectation and thus is more subjective and error-prone than modeling of macromolecules. Poor agreement between calculated and observed density, excessive B factors and the absence of neighbors within non-bonding contact range can help identify modeled waters whose presence is not supported by the data. Though these indicators can change from refinement to refinement as the structure of ordered solvent is varied, larger changes can be expected among spurious waters whose appearance reflects noise in the Fourier-synthesized maps than among solvent whose presence is reliably supported (Pranikar *et al.*, 2009; Terwilliger *et al.*, 2012). Randomly perturb-

ing coordinates, re-refining, and measuring changes in the above statistics provides more reliable detection of outliers than examining the result of a single refinement.

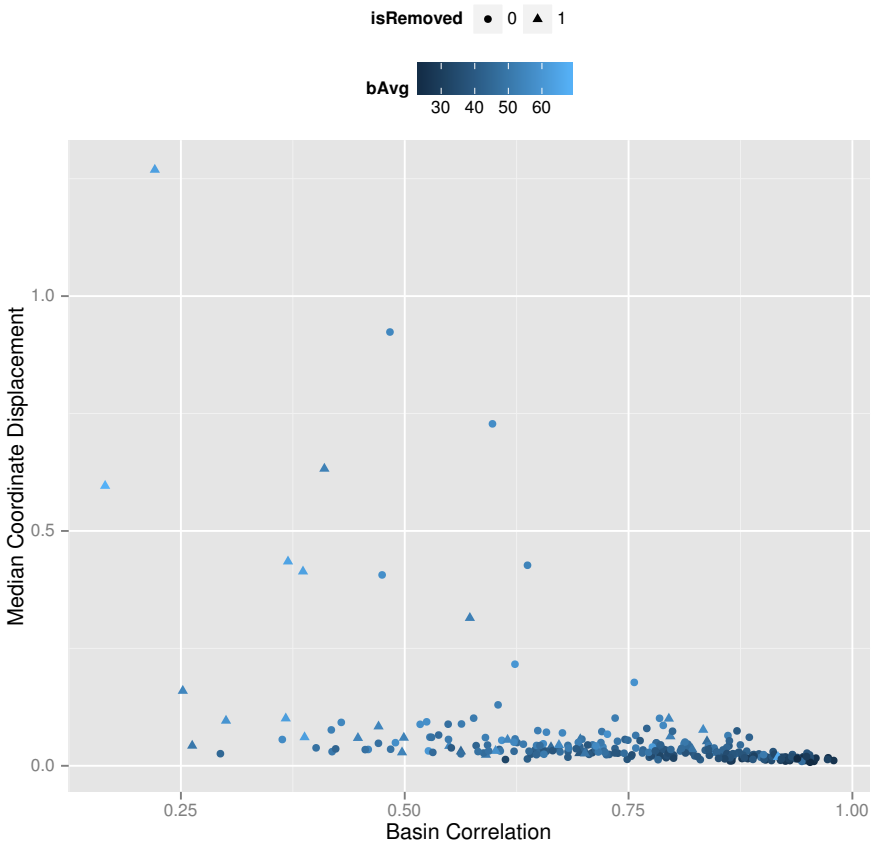


Fig. 7. *Reliability indicators of waters modeled for structure 3ZP8. For 284 waters in structure 3ZP8 assigned to single basins, the correlation between observed and calculated density computed over each basin is plotted against the water's coordinate shift after random perturbation of the model. Color indicates the refined B factor. Deletion of 33 waters marked as triangles yielded negligible change in global R factors.*

To determine whether basin segmentation and calculation of basin properties could help guide solvent modeling, we built a series of alternative 3ZP8 models by deleting subsets of the 289 water molecules included. To help reduce model bias, each model was "shaken" by adding a random displacement selected from a uniform distribution

over the interval $[-0.4, 0.4]\mathring{A}$ and re-refined; this was repeated nine times. The average refined coordinate was calculated and from it the distribution of distances for the nine iterations. Figure 7 shows some metrics for one of the reduced models among 284 of the 289 waters assigned to individual basins. In this case, 33 waters were removed, yielding minor changes in R_{work}/R_{free} from 18.9/21.6 to 18.6/22.3 in the simpler model. For each basin, the median of the distance distribution is plotted against the average real space correlation. Points are colored according to the average value of the isotropic B factor; triangles distinguish waters removed from the original model. It is apparent that the subset of deleted waters, those which contribute little to quality of the model as measured by global R factor statistics, includes many waters with low basin real space correlation and high coordinate displacement. However, this does not hold in all cases. Systematic application of selection criteria based on basin properties for determining a minimal, high-quality, solvent structure requires further investigation.

4. Discussion

We have adapted algorithms for segmenting the $\mathbb{R}^3$ domain of a smooth scalar function on the basis of its gradient to domains that adhere to the symmetry requirements of crystallographic space groups. The hypothesized structural model can be superimposed on the basins that result from this partition.

At sufficiently high resolution, each atom will occupy an individual basin, and, even at more moderate resolution, atoms not constrained by covalent bonds can often be placed in individual basins. This device captures a region of the map within which density is primarily determined by the resident atoms. Various statistics that measure the distribution of density and the agreement between calculated and observed density estimates can be computed over each basin. Basin construction ensures these metrics

reflect primarily the density contributed by the basin's resident atoms. In contrast, regions bounded by surfaces computed from iso-contours of the map density or spheres obtained from element-specific radii provide no assurance that all of an atom's density contribution is being examined.

Segmentation and computation of basin properties can help identify sections of the model with poor fit to the data. Interestingly, larger basins, which encompass regions in which scattering is modeled by bulk solvent, show poorer correlation between calculated and observed density, suggesting this method may help investigate poorly modeled solvent scattering.

Acknowledgments AF acknowledges helpful discussion with Attila Gyulassy, Vijay Nataranjan and other members of the UC Davis Institute for Data Analysis and Visualization. Discussions with Pete Dunten of the Stanford Synchrotron Radiation Lightsource and his review of a draft of this manuscript are gratefully acknowledged. Implementation and validation of our analysis relied on two outstanding libraries for crystallographic computing, *clipper* (Cowtan, 2002) and *cctbx* (Grosse-Kunstleve *et al.*, 2002). We are grateful to the developers for assistance in their use.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Crystallogr D Biol Crystallogr*, **66**(Pt 2), 213–221.

Afonine, P. V. & Urzhumtsev, A. (2004). *Acta Crystallogr A*, **60**(Pt 1), 19–32.

Anderson, M., Schultz, E. P., Martick, M. & Scott, W. G. (2013). *J Mol Biol*.

Bader, R. (1990). *Atoms in Molecules, A Quantum Theory*. Oxford University Press.

Banchoff, T. F. (1970). *The American Mathematical Monthly*, **77**(5), pp. 475–485.

Berman, H., Henrick, K. & Nakamura, H. (2003). *Nature structural biology*, **10**(12), 980.

Collaborative Computational Project, Number 4. 1994, (1994). The ccp4 suite: Programs for protein crystallography.

Cowtan, K. (2002). *The Clipper Project*, vol. 40.

Edelsbrunner, H., Harer, J. & Zomorodian, A. (2003). *Discrete Comput. Geom.* **30**, 87–107.

Forman, R. (1998). *Advances in mathematics*, **134**(1), 90–145.

Grosse-Kunstleve, R. & Bourhis, L. J. (2011). *Computational Crystallography Newsletter*, **2**, 25–28.

Grosse-Kunstleve, R. W., AZwart, P. H., Afonine, Pavel V. andIoerger, T. R. & Adams, P. D. (2006). *Newsletter of the Commission on Crystallographic Computing,International Union of Crystallography*, **5**.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J. Appl. Cryst.* **35**(1), 126–136.

Gyulassy, A., Bremer, P. T., Hamann, B. & Pascucci, V. (2008). *IEEE Trans Vis Comput Graph*, **14**(6), 1619–1626.

Gyulassy, A., Kotava, N., Kim, M., Hansen, C. D., Hagen, H. & Pascucci, V. (2011). *IEEE Trans Vis Comput Graph*.

Gyulassy, A., Natarajan, V., Pascucci, V. & Hamann, B. (2007). *IEEE Trans Vis Comput Graph*, **13**(6), 1440–1447.

Henkelman, G., Arnaldsson, A. & Jónsson, H. (2005). *Comput. Mater. Sci*, **inpress**.

IUCr (1992). *International Tables for Crystallography, Volume C: Mathematical, physical and chemical tables*. International Tables for Crystallography. Dordrecht, Boston, London: Kluwer Academic Publishers, 2nd ed.

King, H., Knudson, K. & Mramor, N. (2005). *Experimental Mathematics*, **14**(4), 435–444.

König, V., Vértesy, L. & Schneider, T. R. (2003). *Acta Crystallogr D Biol Crystallogr*, **59**(Pt 10), 1737–1743.

Malcolm, N. O. J. & Popelier, P. L. A. (2003). *Journal of Computational Chemistry*, **24**(10), 1276–1282.

Marwan, N., Kurths, J., Thomsen, J. S., Felsenberg, D. & Saparin, P. (2009). *Phys. Rev. E*, **79**, 021903.

Matsumoto, Y. (1997). *An Introduction to Morse Theory*. The American Mathematical Society.

Muller, P. (2006). *Crystal Structure Refinement: A Crystallographer's Guide to SHELXL*. IUCr texts on crystallography. OUP Oxford.

Pranikar, J., Afonine, P. V., Guncar, G., Adams, P. D. & Turk, D. (2009). *Acta Crystallogr D Biol Crystallogr*, **65**(Pt 9), 921–931.

Robins, V., Wood, P. J. & Sheppard, A. P. (2011). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **33**(8), 1646–1658.

Roerdink, J. B. & Meijster, A. (2000). *Fundam. Inf.* **41**(1,2), 187–228.

Rupp, B. (2006). *Nature*, **444**(7121), 817–817.

Sheldrick, G. M. (2008). *Acta Crystallogr A*, **64**(Pt 1), 112–122.

Tang, W., Sanville, E. & Henkelman, G. (2009). *Journal of Physics: Condensed Matter*, **21**(8), 084204.

Terwilliger, T. C., Read, R. J., Adams, P. D., Brunger, A. T., Afonine, P. V., Grosse-Kunstleve, R. W. & Hung, L. W. (2012). *Acta Crystallogr D Biol Crystallogr*, **68**(Pt 7), 861–870.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta crystallographica. Section D, Biological crystallography*, **67**(Pt 4), 235–242.

Zhang, K. Y. J., Cowtan, K. D. & Main, P. (2001). In *International Tables for For Crystallography Volume F*, edited by M. Rossmann & E. Arnold. Kluwer Academic Publishers.

Zomorodian, A. (2009). In *Algorithms and Theory of Computation Handbook*, edited by M. J. Atallah & M. Blanton, vol. 2. CRC Press.

Synopsis