TEXTAL:
A Pattern Recognition System for Interpreting Electron Density Maps

Thomas R. Ioerger
Department of Computer Science

James C. Sacchettini
Center for Structural Biology

Texas A&M University

Other Contributors:
Tom Holton (MS student)
Jon Christopher (post-doc)

Funding Acknowledgements:
National Institutes of Health
Texas A&M University

Pattern Recognition

Look for regions with similar patterns of electron density in previously solved maps

Treat map as overlapping spheres (e.g. 5Å radius)

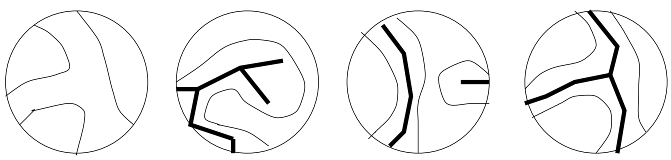
Database search

Create maps from structure factors (e.g. 3Å res.)

Parse into spherical regions, with local coordinates

Find "best match"

Rotate/translate local coords into new position



Learning from examples

- Exploits crystallographic databases, increasing size
- Can have examples of atypical patterns too (e.g. Gly)

Defining Similarity of Density Patterns

Density correlation: similarity between density patterns in two regions (summed over lattice points in region Ω)

$$corr(R_1, R_2) = \frac{\sum_{\Omega} (\rho_1 - \bar{\rho_1})(\rho_2 - \bar{\rho_2})}{\sqrt{\sum_{\Omega} (\rho_1 - \bar{\rho_1})^2 \sum_{\Omega} (\rho_2 - \bar{\rho_2})^2}}$$

Maximize over all orientations: *rotation search*

- exhaustive search (slow but accurate)
- align moments of inertia (fast but less reliable)
- peak-matching (heuristic, good compromise)

Feature Extraction

Improving efficiency of database search

Density corr is too slow to run on 1000's of regions

Extract numeric features that capture aspects of density patterns

Use features for fast matching (lookup)

Must be *rotation invariant* - value constant regardless of orientation

Features

- average density in region

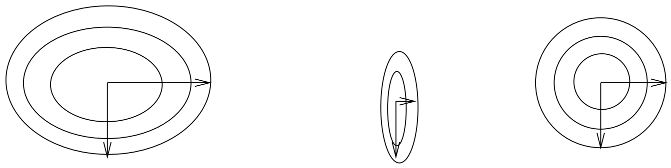
$$\textit{Avg Dens} = \frac{1}{n}\sum_i \rho_i$$

- distance to center of mass

$$\textit{Cen of Mass} = \langle c_x, c_y, c_z \rangle = \langle \frac{\sum \rho_i x_i}{\mu}, \frac{\sum \rho_i y_i}{\mu}, \frac{\sum \rho_i z_i}{\mu} \rangle$$

$$\textit{Dist} = \sqrt{(c_x - c_0)^2 + (c_y - c_0)^2 + (c_z - c_0)^2}$$

- moments of intertia (magnitudes and ratios)

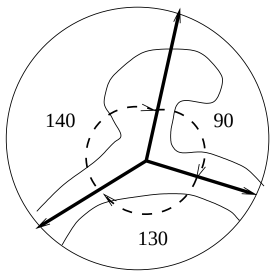


- statistical features (std. dev., kurtosis...)

$$\mathit{Std\ Dev} = \frac{1}{n}\sum_i(\rho_i - \mu)^2$$

$$\mathit{Kurtosis} = \frac{1}{n}\sum_i(\rho_i - \mu)^4/\sigma^4$$

- geometric features - shapes (e.g. spoke angles)



- contours: volume, surface area, connectivity...

Features can be calculated over different radii ($3-6\mathring{A}$)

Feature-based similarity metric

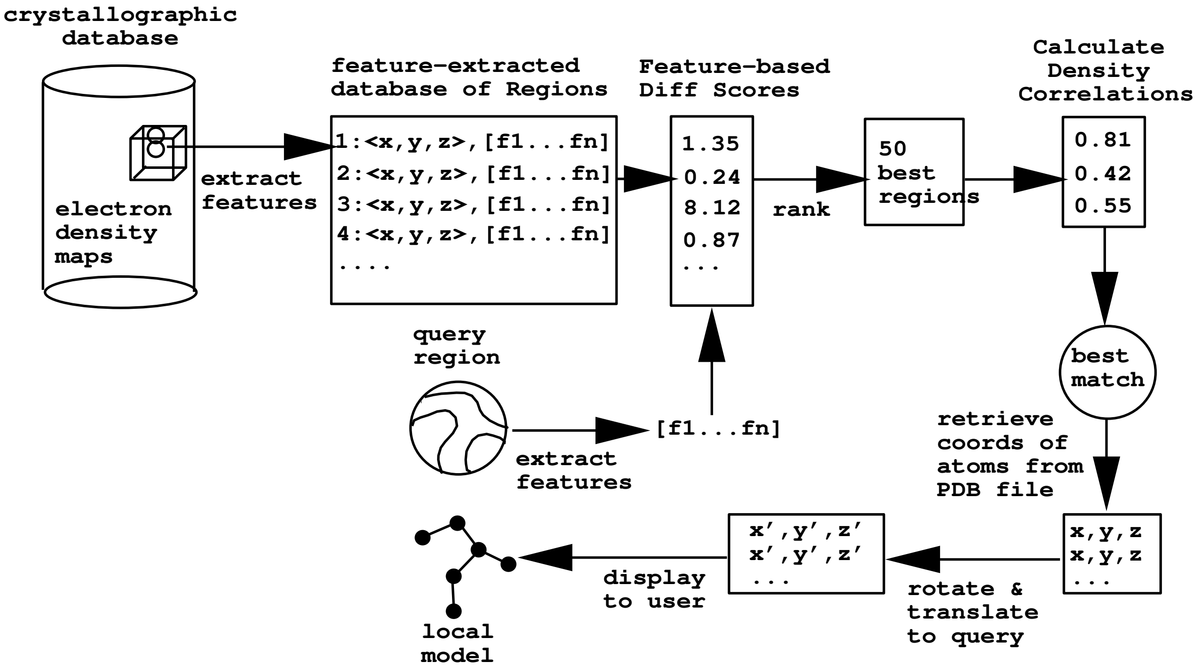
weighted Euclidean distance metric

$$\textit{diff}(R_1, R_2) = \sqrt{\sum_{i=1}^{n} \; w_i(F_i(R_1) - F_i(R_2))^2}$$

Optimization of feature weights

- Irrelevant features can add noise
- Uniform mixture with normalized features
- Slider Algorithm - optimizes ranking of good matches (high dens. corr.) against poor matches
- Other Methods - feature correlations, wrapper methods, RELIEF

TEXTAL Architecture



Two phases:

1. Retrieve top K=50 regions with minimum feature distance (some spurious matches)

2. Calculate density correlations by rotation search; Predict coords based on region with highest corr.

Results

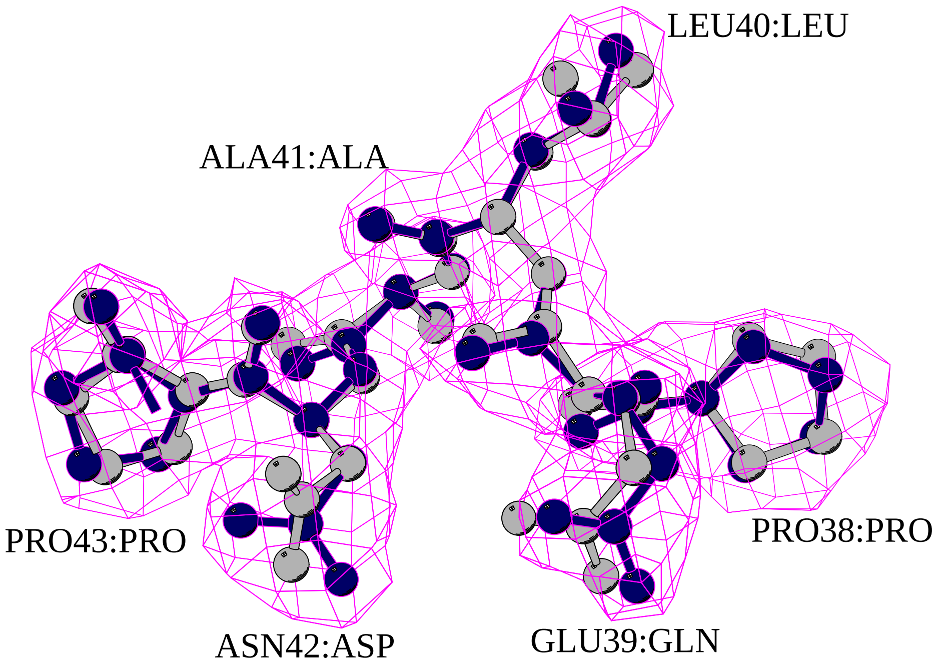
Experiment 1: Modeling Gaussian map for 1udi

- cover structure with density dispersed around atoms
- run TEXTAL on each C_α atom (assume known)
- average density correlation of matches: 0.70
- returned residues:
  - 42% identical
  - 61% structurally similar
  - 5% backbones flipped

Experiment 2: Modeling real maps created from structure factors (F_{obs})

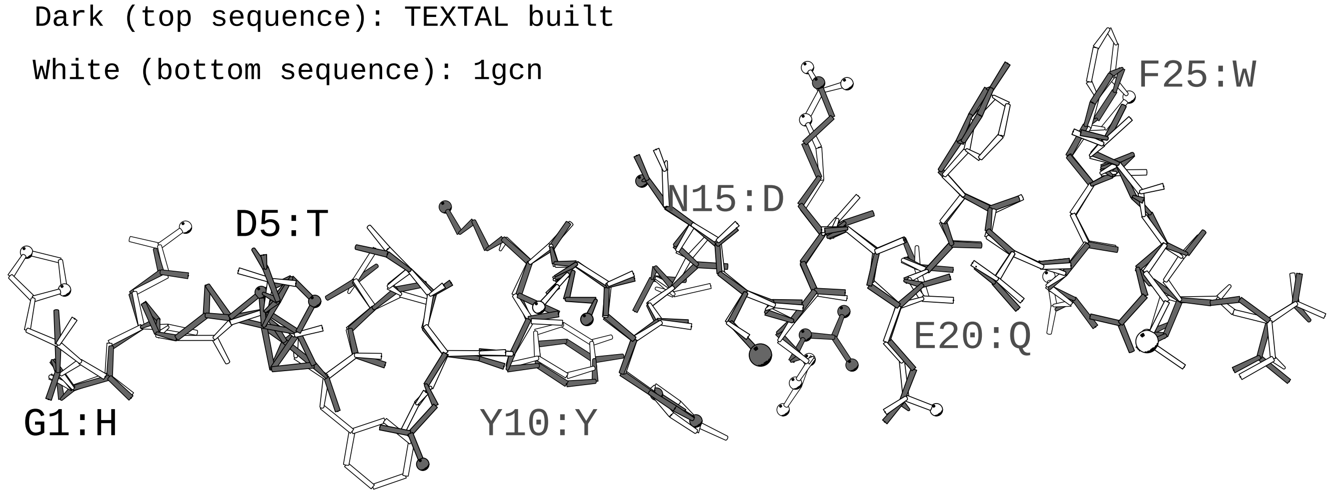
- phases calculated from structures: ϕ_{calc}
- crambin (1ab1), flavodoxin (1ag9) - 2.8Å res.
- database: 50 maps back-transformed from structures (with scattering by temp factors): F_{calc}

method	protein	all-atom RMS	main-chain RMS	side-chain RMS
TEXTAL	crambin	0.92	0.75	1.25
	flavodoxin	0.95	0.74	1.28
SMM (Levitt)	crambin	1.51	0.64	2.17
	flavodoxin	1.71	0.54	2.39
MaxSprout (Holm&Sander)	crambin			2.12
	flavodoxin	1.57	0.48	2.19



Dark (top sequence): TEXTAL built

White (bottom sequence): 1gcn



G S D G D H T S Q Y K K H L N C R K E E L Y T L F L E I T
H S Q G T F T S D Y S K Y L D S R R A Q D F V Q W L M N T
1 5 10 15 20 25

CAPRA:
C-Alpha Pattern Recognition Algorithm

C_α's needed for modeling in TEXTAL (acc. to $< 1\mathring{A}$)

Not identified uniquely in other main-chain tracing programs (e.g. Bones)

Use pattern recognition again

1. Create database with spheres centered on lattice points $0 - 5\mathring{A}$ away from a true C_α in known maps
2. For each lattice point in contour of unknown map, look up most similar regions and estimate distance
3. Do some clustering to predict C_α's