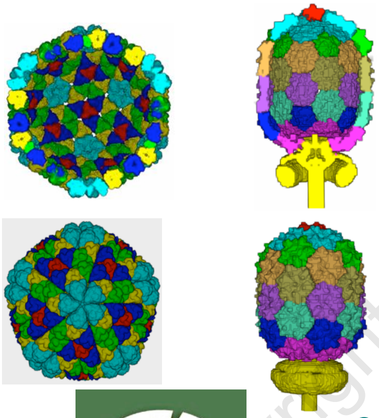
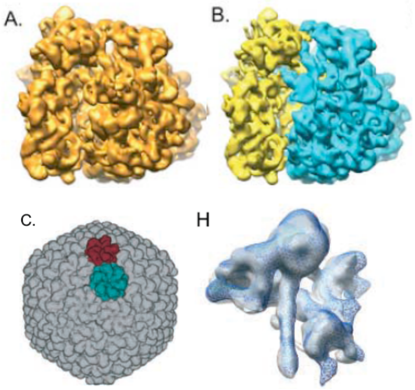
Molecular Structure and Properties Elucidation from 3D Electron Microscopy

Chandrajit Bajaj



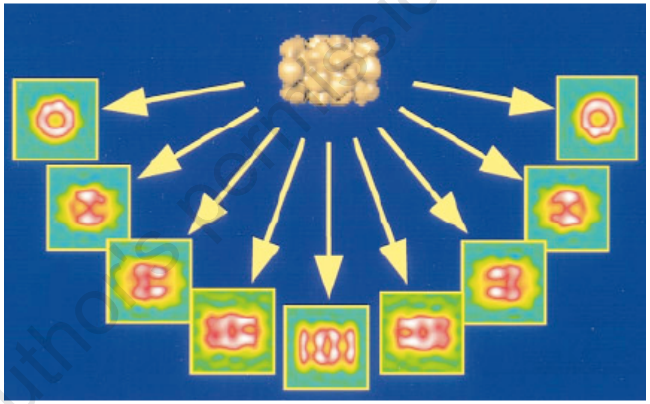


Center for Computational Visualization http://www.ices.utexas.edu/CCV
Institute of Computational and Engineering Sciences
Department of Computer Sciences
University of Texas at Austin

May 2007

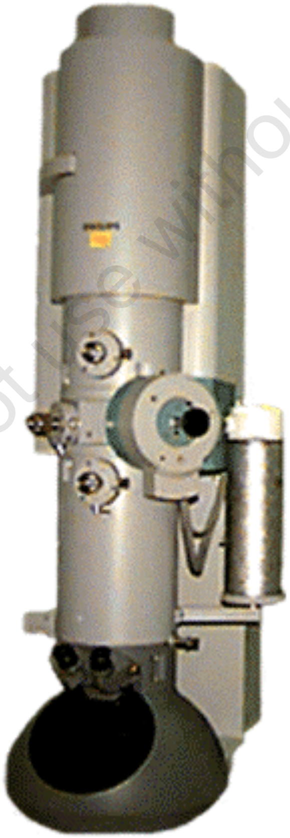
Multi-scale Multi-Modal Imaging

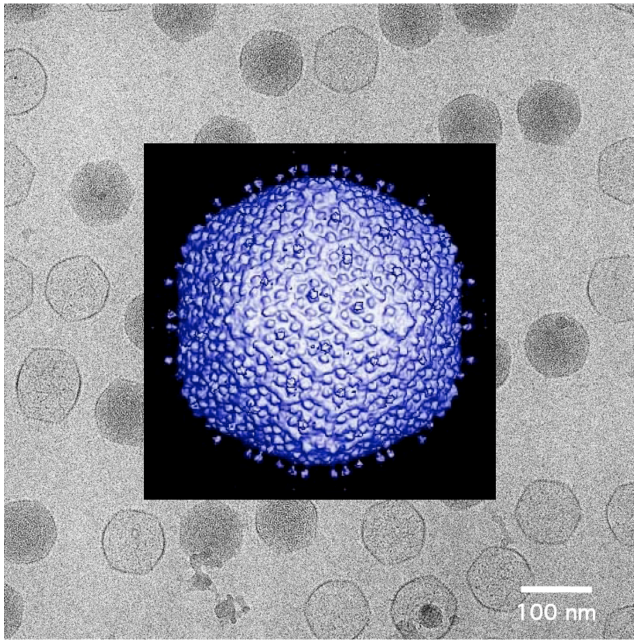
- X-ray crystallography (diffraction)
  - Atomic resolution
  - Difficulties (experimental, computational)
- Nuclear magnetic resonance (NMR)
  - Atomic resolution
  - Limited to small structures
- Electron Microscopy
  - Cryo-electron tomography
    - Low resolution (20Å - 200Å)
    - Good for whole cell or cell organelles
  - Single particle cryo-EM
    - Intermediate resolution (5Å - 20Å)
    - Computationally more demanding



Electron tomography

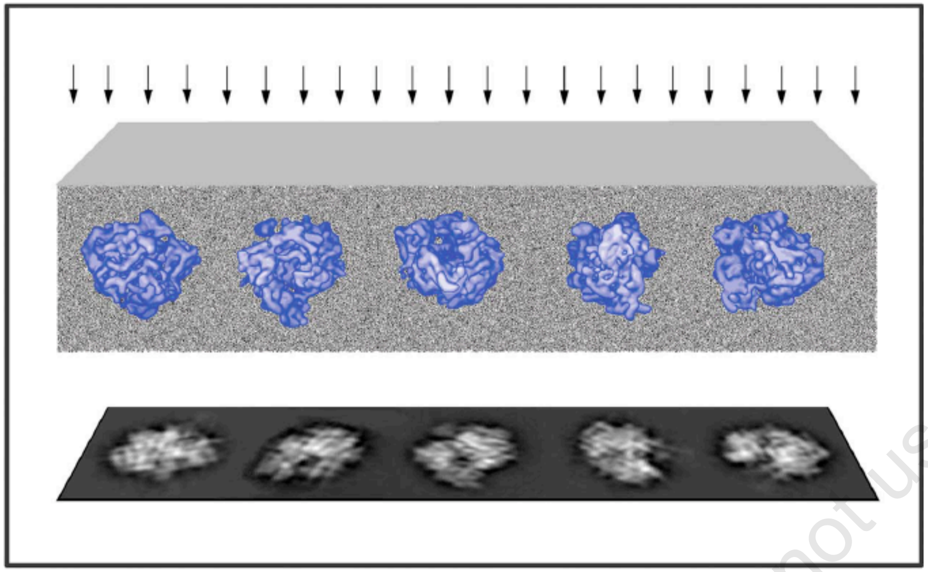
(Picture from A.J. Koster et al, JSB, 1997)



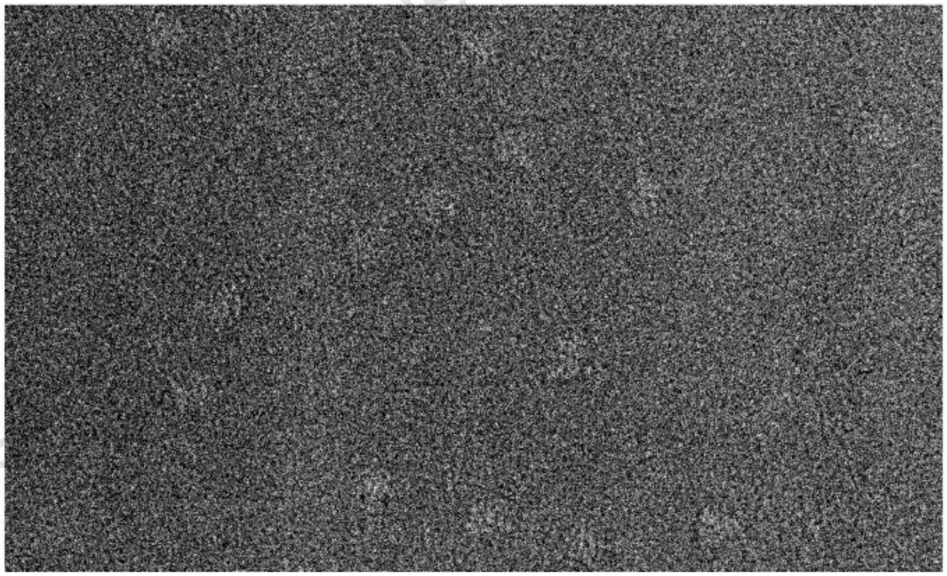


Single particle cryo-EM

Image Acquisition



schematic of possible orientations

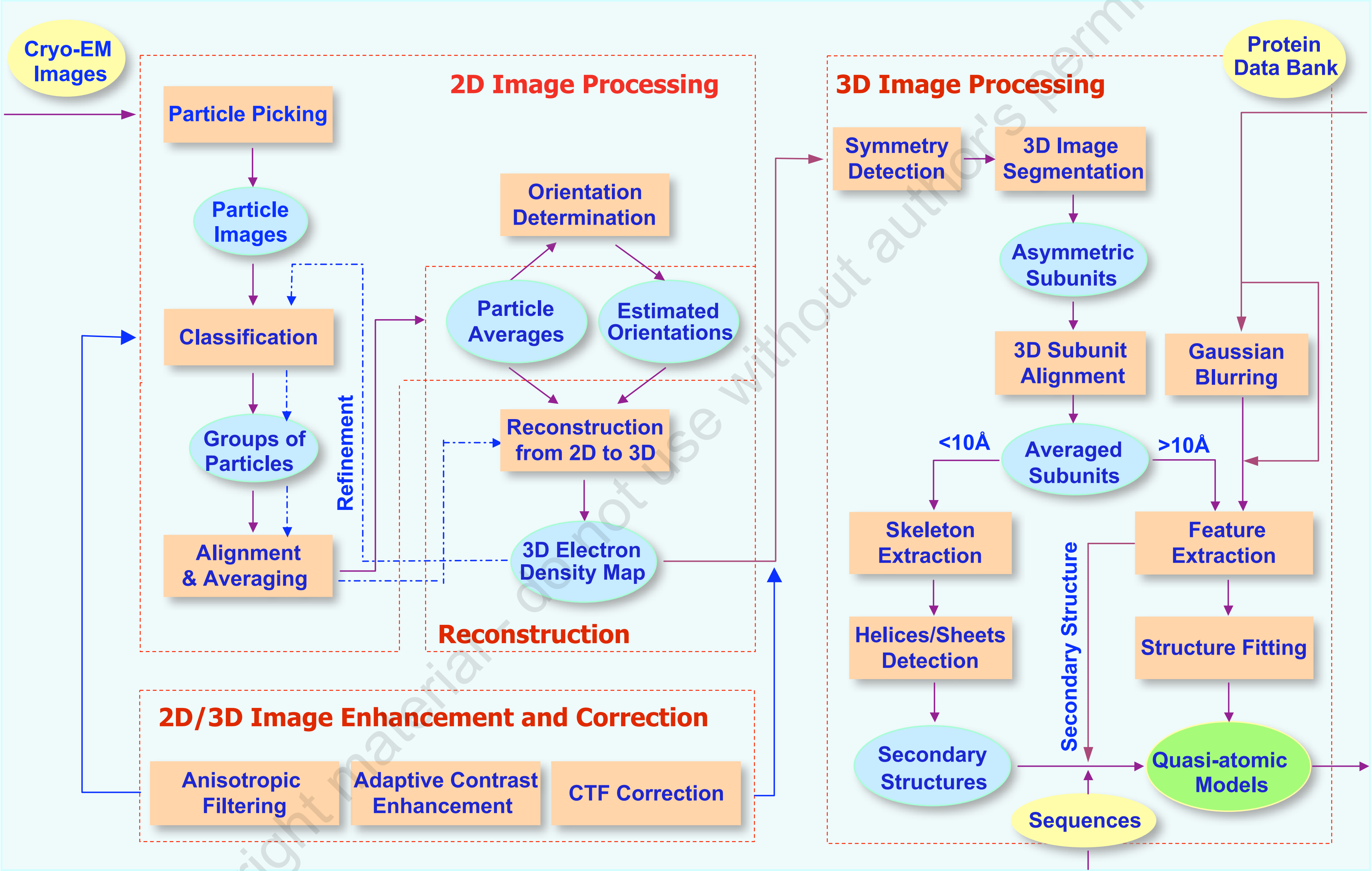


2D Cryo-EM map

- Images collected via electron microscopy
- Lowest possible radiation used to limit damage to sample

Annu. Rev. Biophys. Biomol. Struct. 2002. 31:303–19

Quasi Atomic Models from Single Particle Cryo-EM



Cryo-EM Images

2D Image Processing

Particle Picking

Particle Images

Classification

Groups of Particles

Alignment & Averaging

Refinement

Orientation Determination

Particle Averages

Estimated Orientations

Reconstruction from 2D to 3D

3D Electron Density Map

Reconstruction

2D/3D Image Enhancement and Correction

Anisotropic Filtering

Adaptive Contrast Enhancement

CTF Correction

3D Image Processing

Protein Data Bank

Symmetry Detection

3D Image Segmentation

Asymmetric Subunits

3D Subunit Alignment

Gaussian Blurring

Averaged Subunits

<10Å

>10Å

Skeleton Extraction

Helices/Sheets Detection

Secondary Structures

Feature Extraction

Structure Fitting

Secondary Structure

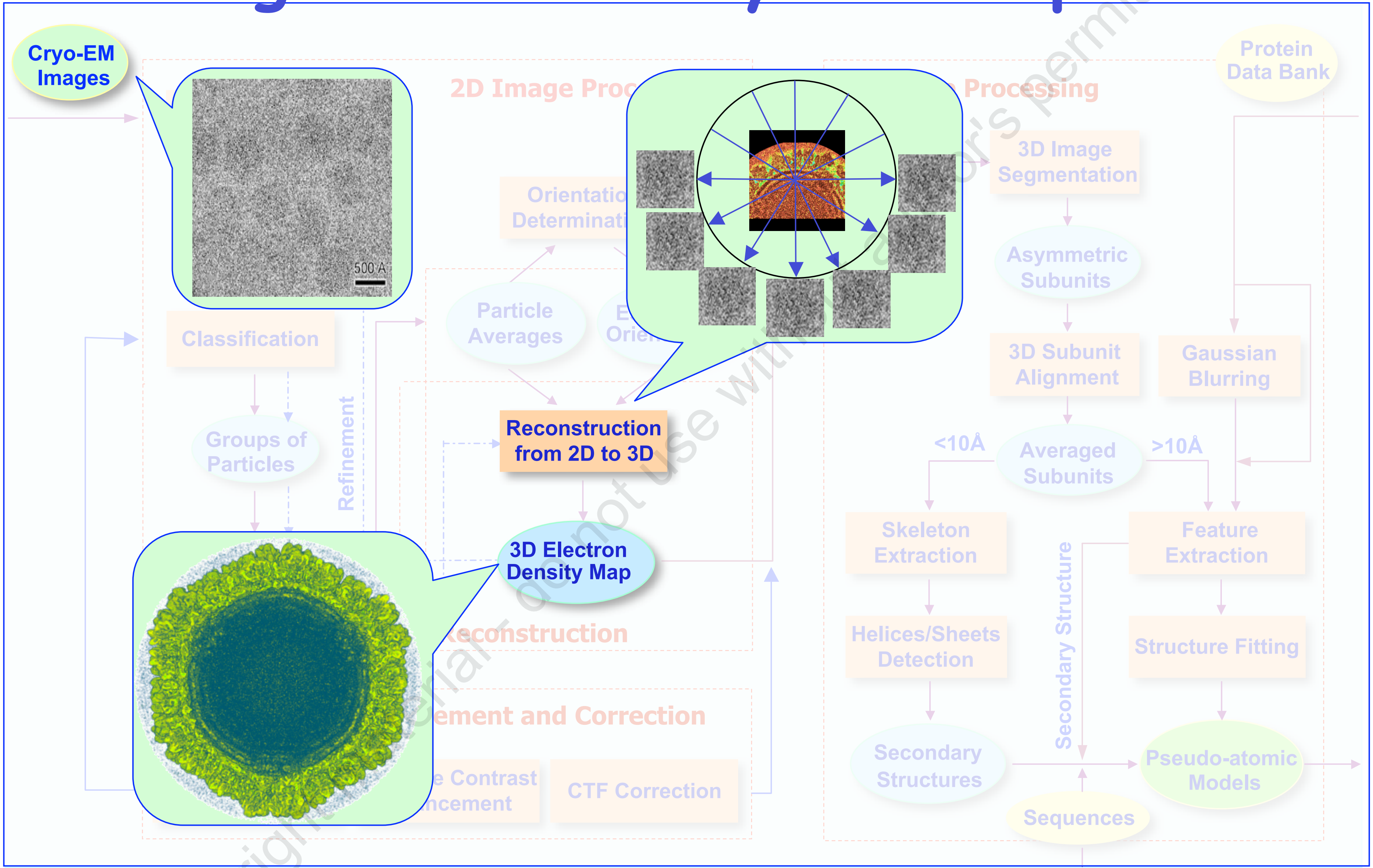
Sequences

Quasi-atomic Models

Center for Computational Visualization http://www.ices.utexas.edu/CCV
Institute of Computational and Engineering Sciences
Department of Computer Sciences
University of Texas at Austin

May 2007

Single Particle Cryo-EM Pipeline



Cryo-EM Images

500 Å

2D Image Proc

Orientatio Determinati

Particle Averages

E Orien

Classification

Groups of Particles

Refinement

Reconstruction from 2D to 3D

3D Electron Density Map

econstruction

ement and Correction

e Contrast ncement

CTF Correction

Processing

Protein Data Bank

3D Image Segmentation

Asymmetric Subunits

3D Subunit Alignment

Gaussian Blurring

<10Å

Averaged Subunits

>10Å

Skeleton Extraction

Feature Extraction

Helices/Sheets Detection

Secondary Structure

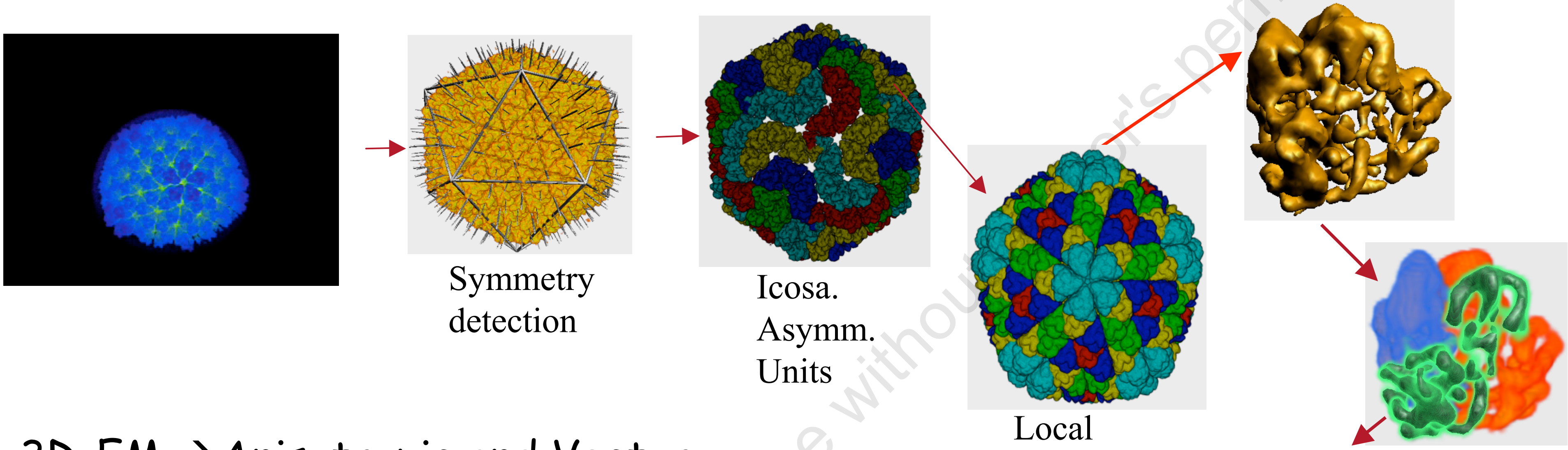
Structure Fitting

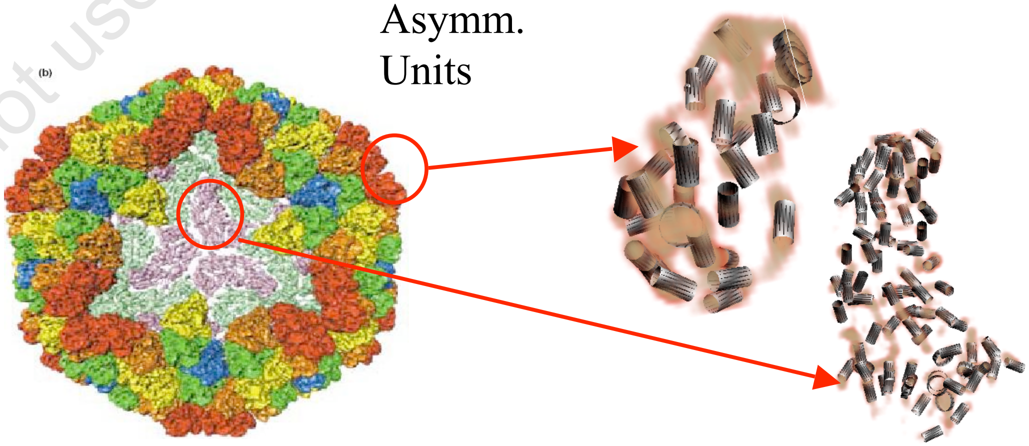
Secondary Structures

Pseudo-atomic Models

Sequences

Sub-nanometer Structure Elucidation from 3D-EM





3D-EM →Anisotropic and Vector Diffusion Filtering → Symmetry Detection --> Structure Segmentation → Secondary/Tertiary Structure →Match & Fit --> Quasi-Atomic Models

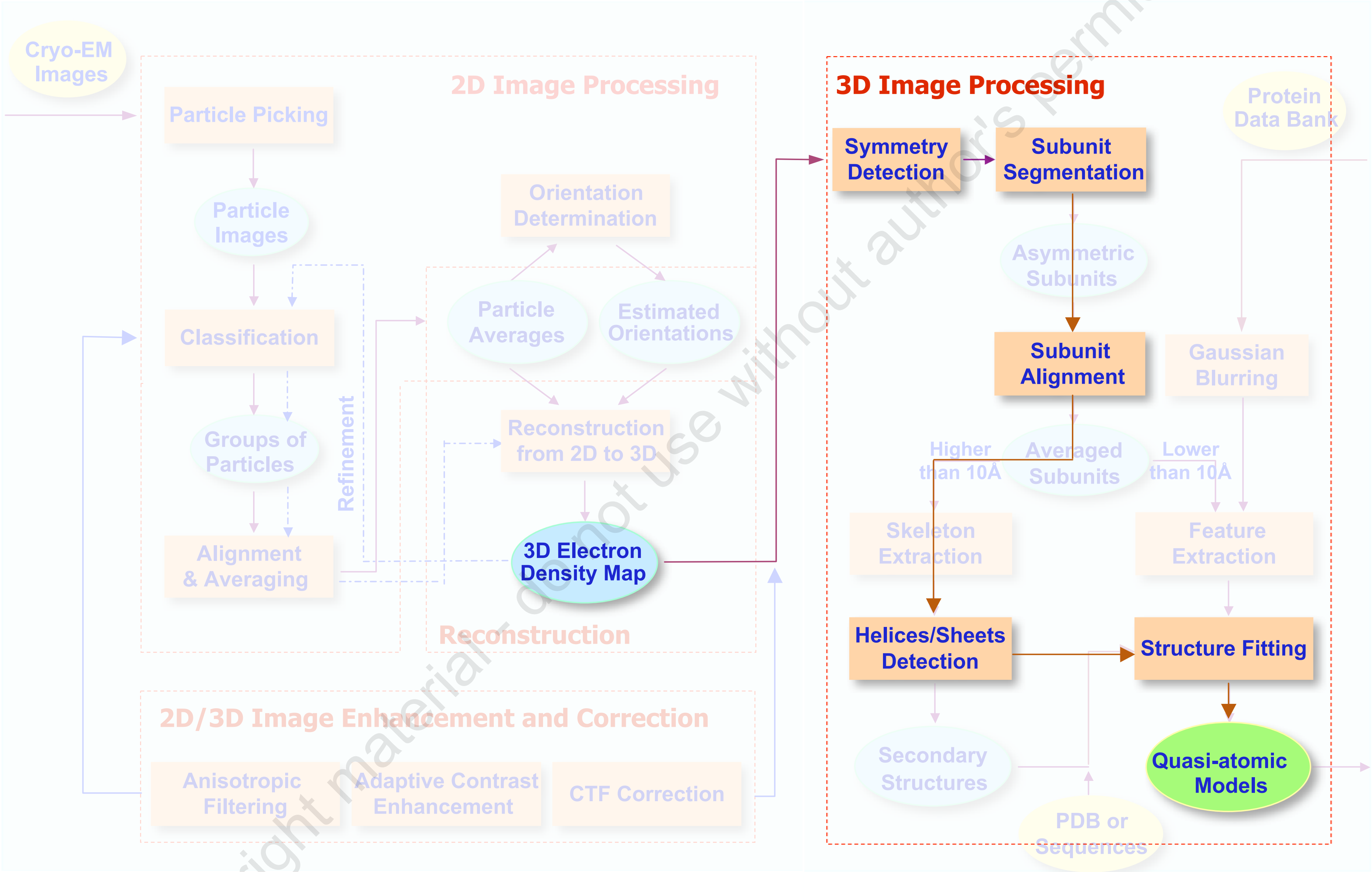
C. Bajaj, Z. Yu, Geometric and Signal Processing for 3DEM , Handbook in Comp. Biology, 2005, ed. S. Aluru

Rice Dwarf Virus

(CryoEM Data courtesyrs: Dr. Wah Chiu (Baylor),)

** NSF-ITR, NIH

Single Particle Cryo-EM: Automatic Structure Analysis



Cryo-EM Images

2D Image Processing

Particle Picking

Particle Images

Classification

Groups of Particles

Alignment & Averaging

Refinement

Orientation Determination

Particle Averages

Estimated Orientations

Reconstruction from 2D to 3D

3D Electron Density Map

Reconstruction

2D/3D Image Enhancement and Correction

Anisotropic Filtering

Adaptive Contrast Enhancement

CTF Correction

3D Image Processing

Symmetry Detection

Subunit Segmentation

Protein Data Bank

Asymmetric Subunits

Subunit Alignment

Gaussian Blurring

Higher than 10Å

Averaged Subunits

Lower than 10Å

Skeleton Extraction

Feature Extraction

Helices/Sheets Detection

Structure Fitting

Secondary Structures

Quasi-atomic Models

PDB or Sequences

Center for Computational Visualization http://www.ices.utexas.edu/CCV
Institute of Computational and Engineering Sciences
Department of Computer Sciences
University of Texas at Austin

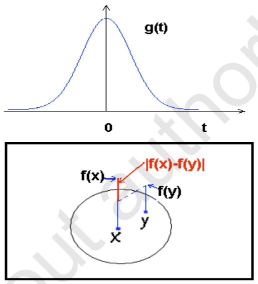
May 2007

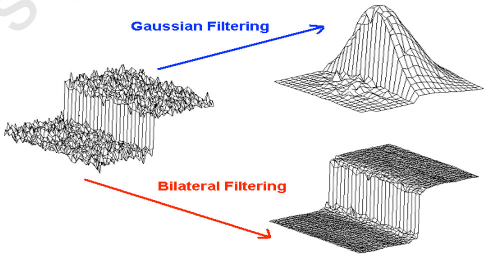
Step #1 : Anisotropic Filtering

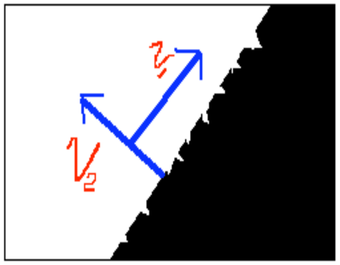
Bilateral filtering

$$h(x,\xi) = e^{-\frac{(x-\xi)^2}{2\sigma_d^2}} \cdot e^{-\frac{(f(x)-f(\xi))^2}{2\sigma_r^2}}$$

where σ_d and σ_r are parameters and $f(.)$ is the image intensity value.





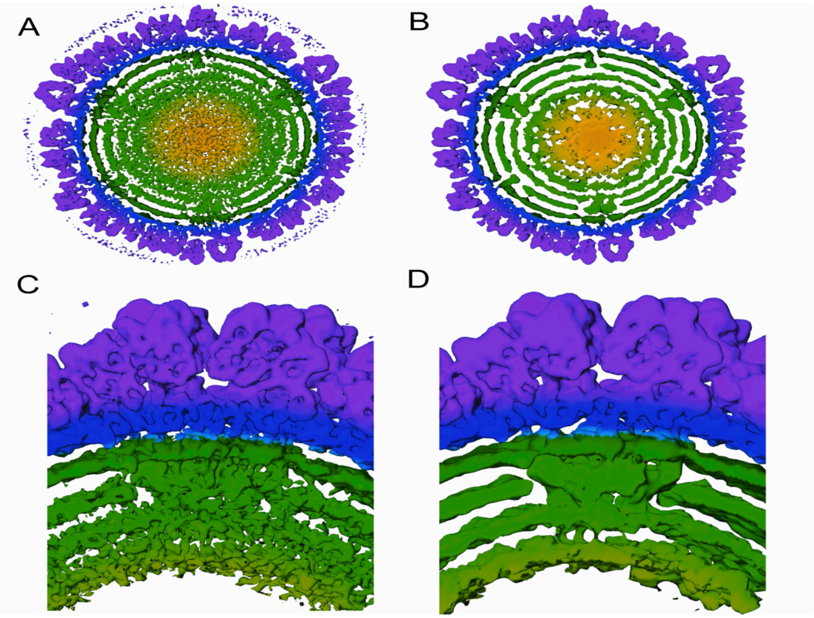


Anisotropic diffusion filtering

$$\partial_t \phi - \mathrm{div}(a(|\nabla \phi_\sigma|)\nabla \phi) = 0$$

where **a** stands for the diffusion tensor determined by local curvature estimation.

C. Bajaj, G. Xu, ACM Transactions on Graphics, (2003),22(1), pp. 4- 32.



W. Jiang, M. Baker, Q. Wu, C. Bajaj, W. Chiu, Journal of Structural Biology, 144, 5,(2003), Pages 114-122

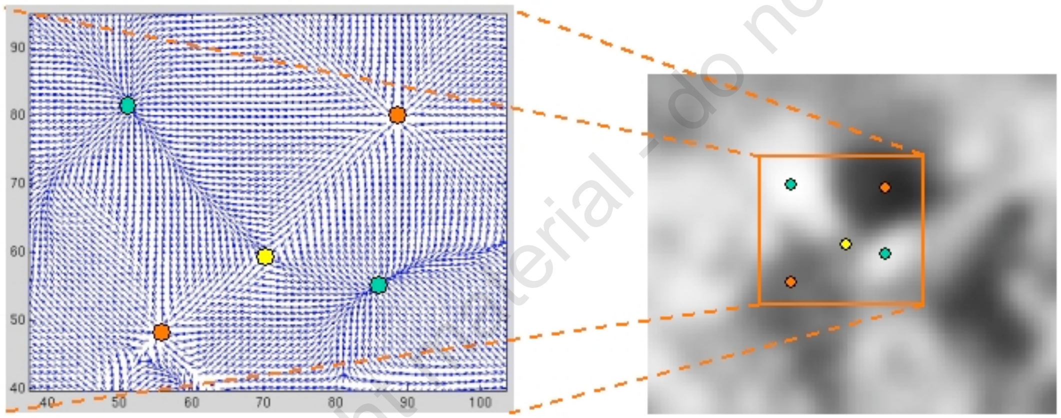
Step #2: Critical Point Detection

- For smooth data:
  - zeroes of the gradient vector field
  - simple, easy to implement
- For noisy data:
  - Gradient vector diffusion
  - higher time complexity but robust to noise

- Gradient vector diffusion:
  - smoothing the vector fields
  - diffusion to flat regions

$$\begin{cases} \dfrac{\partial u}{\partial t} = \mu \cdot div(g(\alpha)\nabla u) \\ \dfrac{\partial v}{\partial t} = \mu \cdot div(g(\alpha)\nabla v) \end{cases}$$

where $g(\alpha)$ is a decreasing function α is the angle between the central pixel and its surrounding pixels.



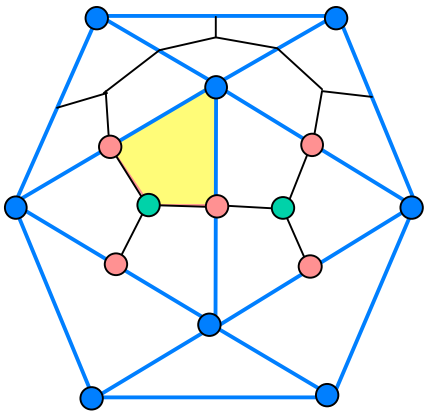
● minimum (0)

● maximum (3)

● saddle (1, 2)

Step #3: Symmetry Detection

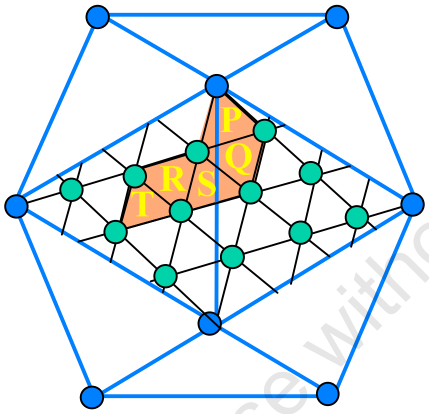
- Asymmetric subunits in an icosahedra



● Two-fold vertices

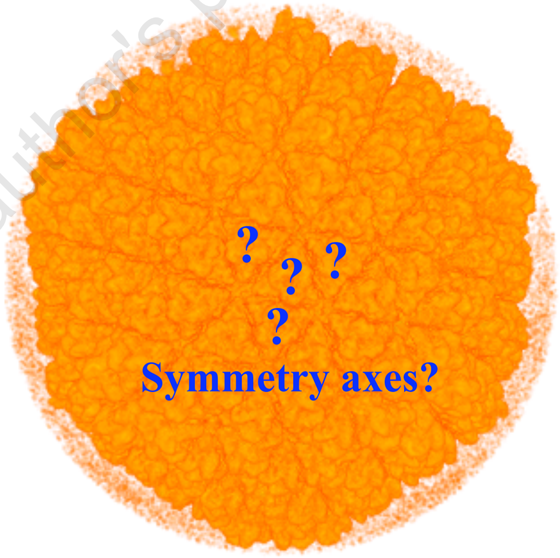
● Three-fold vertices

● Five-fold vertices



Local symmetry (RDV)

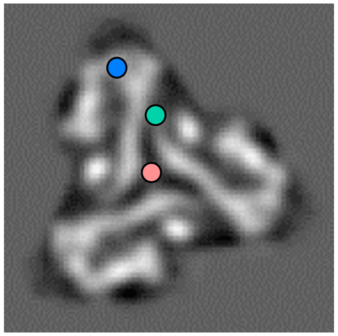
(260 trimers or 720 proteins)



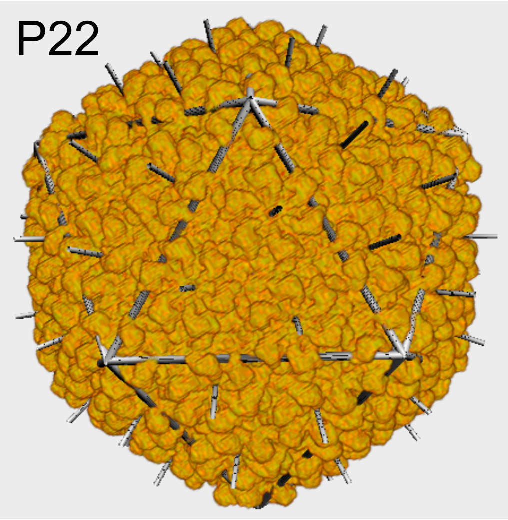
- Correlation search, addtly sped up by Spherical FFT:

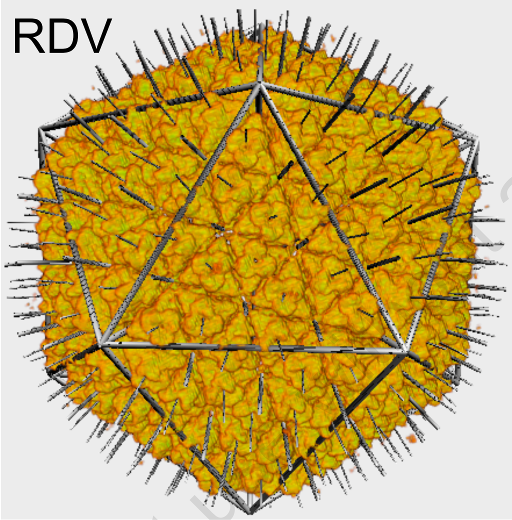
Find best c, minimizing: $$\sum_{\vec{r}\in D_0}\left(f(\vec{r}) - f(R_{2\pi/n}(c)\cdot\vec{r})\right)^2$$

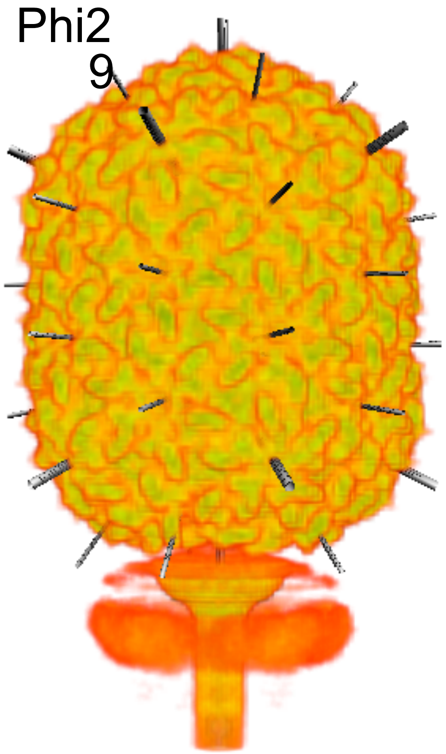
Y.Zu, C. BajajIEEE Transactions on Image Processin, 2005, 14, 9, 1324-1337

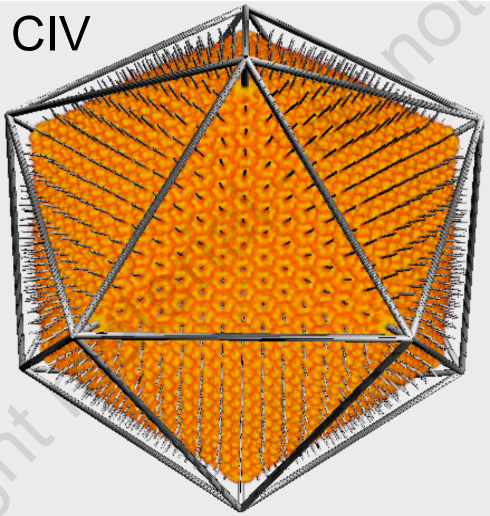


Results of Automatic Symmetry Detection in Virus and Phage Capsid Shells







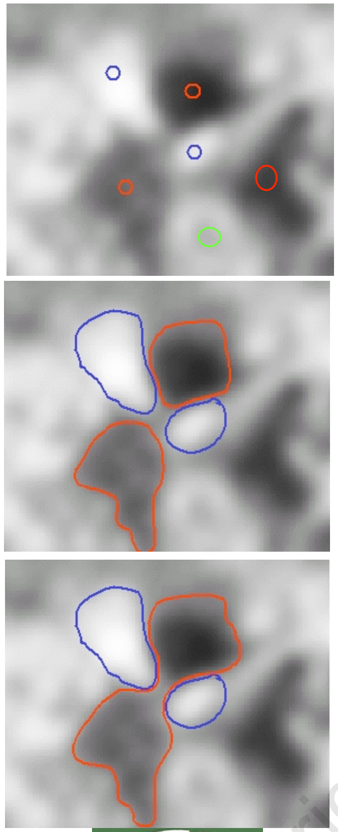


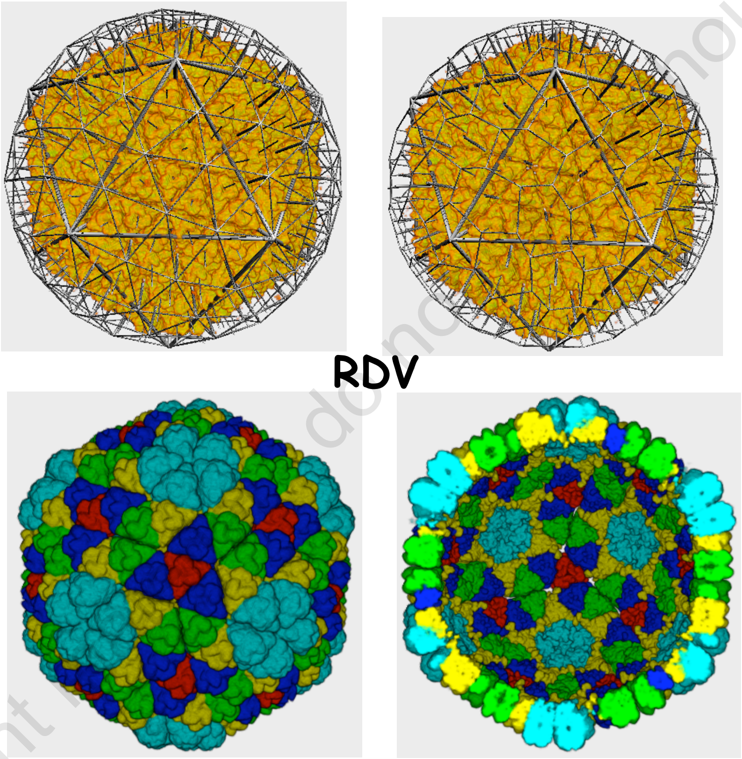
Data Courtesy: Wah Chiu (BCM), Tim Baker (UCSD)

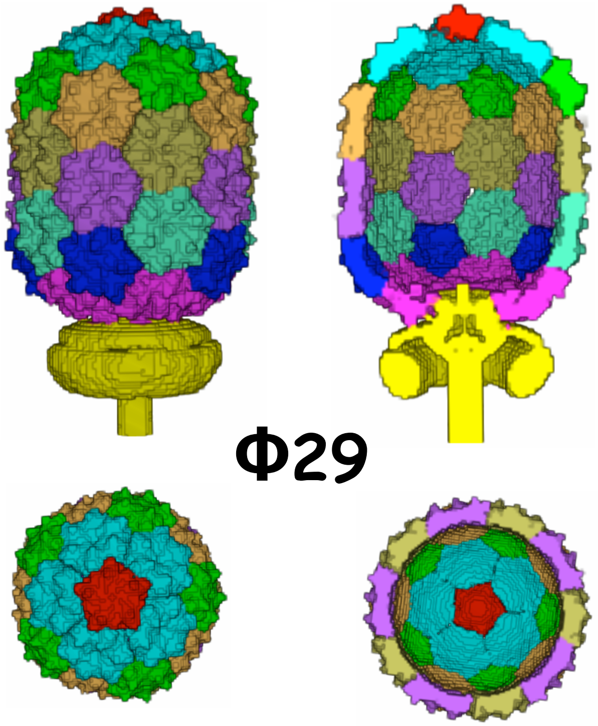
Step #4: Subunit Segmentation

- Multi-seed Fast Marching Method
  - Classify the critical points based on local symmetry into separate groups.
  - Each seed initializes one contour, with its group's membership.
  - Contours march simultaneously. Contours with same membership are merged, while contours with different membership stop each other.

Zeyun, Bajaj IEEE Trans on Imag. Proc.,2005



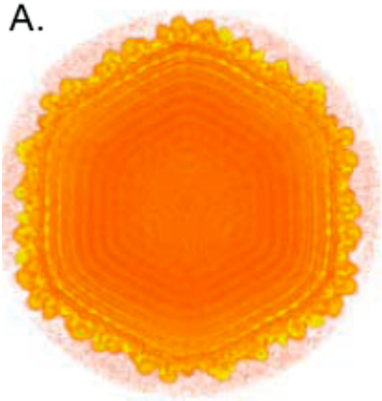




Φ29 data courtesy: Tim Baker (UCSD)

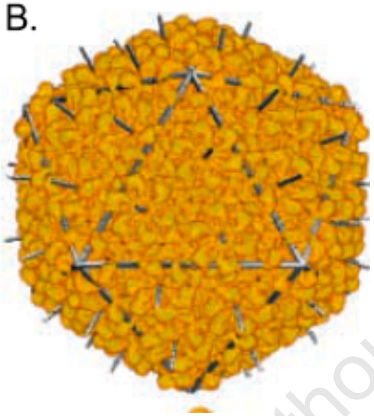
Bacteriophage P22

A.



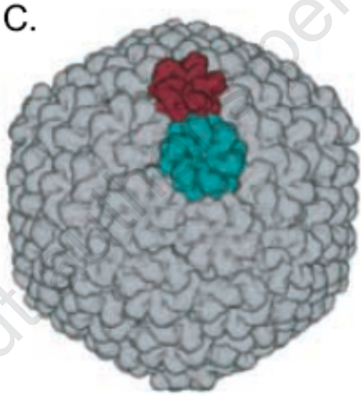
9.5 Angstrom Density Map

B.



Global Icosahedral Symmetry

C.



Local 6-fold Symmetry

D.

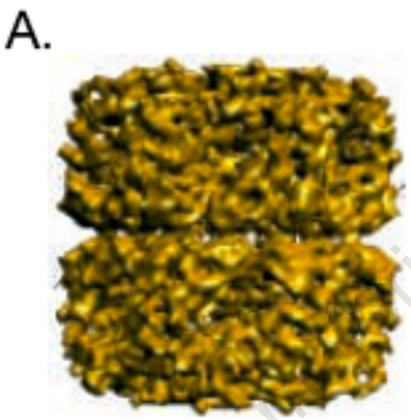
Subunit about global Icosahedral 5 fold axis

F.

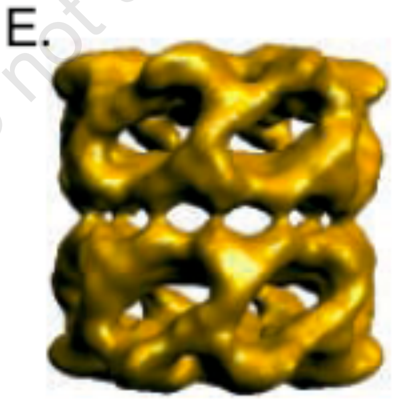
Subunit about local 6 fold axis

GroEL

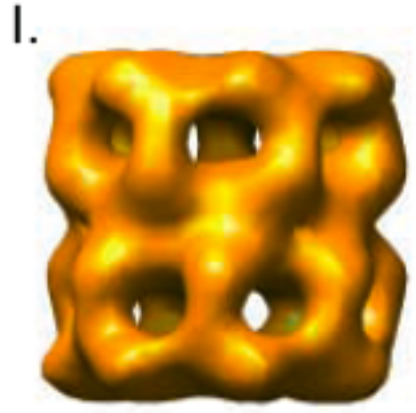
- Chaperonin responsible for protein folding
- Composed of 14 identical subunits
- Subunits have D7 symmetry



6 angstrom

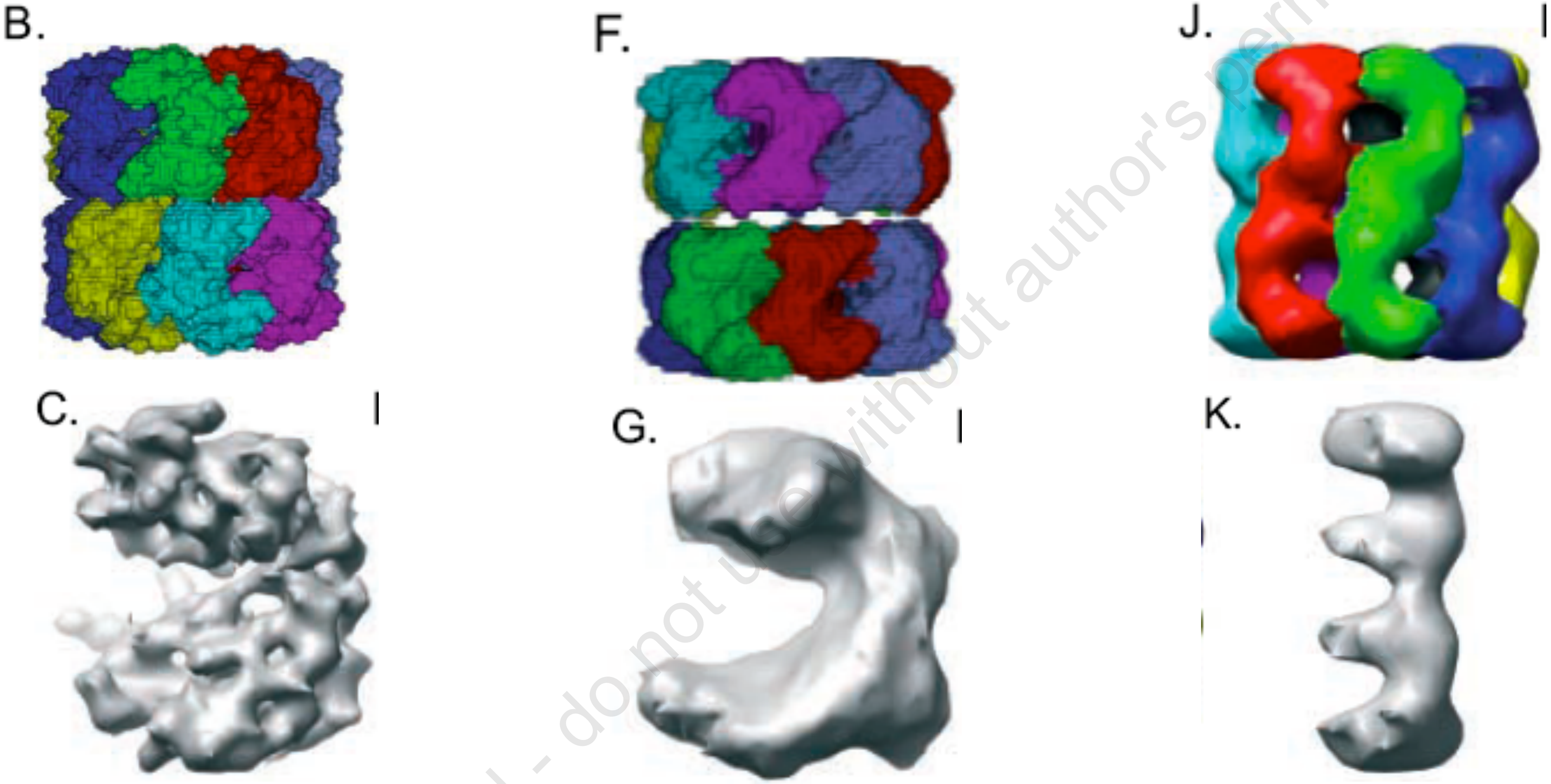


11.5 angstrom



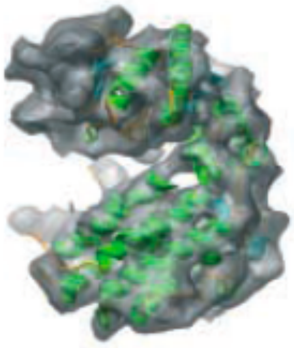
25 angstrom

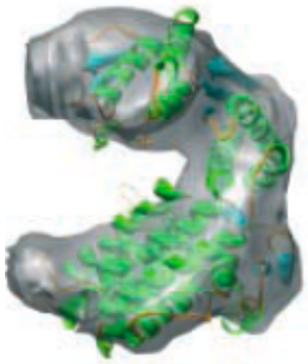
GroEL

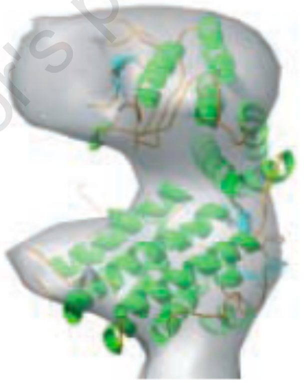


- Segmentation from 6 angstrom and and 11.5 angstrom Cryo-EM structures yielded 14 monomers as expected
- Segmentation of 25 angstrom Cryo-EM structure yielded seven segments, each composed of 2 monomers

GroEL Correlation Values


6 Angstrom
.76 correlation

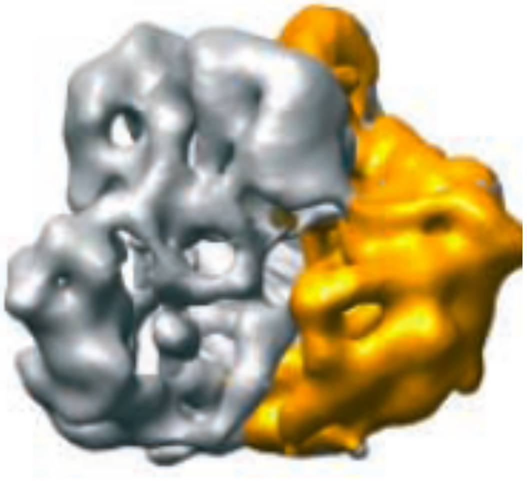

11.5 Angstrom
.75 correlation


25 Angstrom
.52 correlation

- Gray= automated segmentation
- Green= crystal structure
- Correlation value generated against x-ray structure
- Accuracy of segmentation relies on resolution of Cryo-EM map
  - –Segments don’t always correspond to subunits from low resolution maps

Additional Results

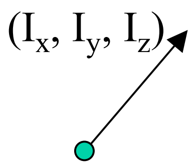
Subunit	Model	Correlation
RDV P8 trimer	Manual segmentation, P8 trimer	0.74
RDV P8 trimer	X-ray structure, P8 trimer	0.85
RDV P8 monomer	Manual segmentation, P8 monomer	0.80
RDV P8 monomer	X-ray structure, P8 monomer	0.84
P22 tail machine	X-ray structure, P22 trimeric tail spike	0.76
70S ribosome, 50S subunit	X-ray structure, 50S subunit	0.63
70S ribosome, 30S subunit	X-ray structure, 30S subunit	0.73
30S subunit, RNA	X-ray structure, 30S subunit RNA	0.66



M. Baker, Z. Yu, W. chiu, C. Bajaj, J of Structural Biology, 2006

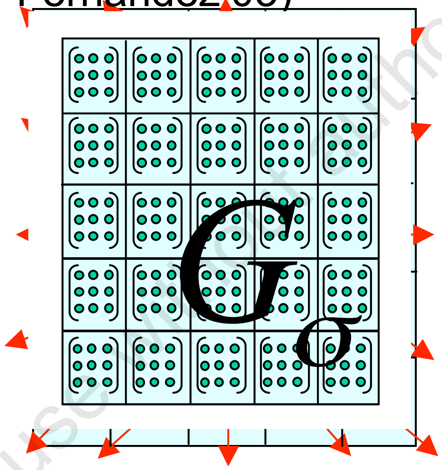
Step #5A: Secondary Structure Identification

- Gradient tensor



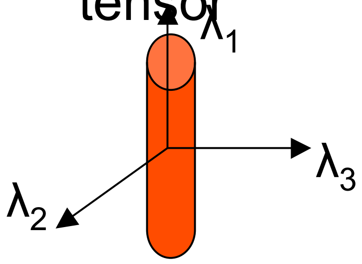
$$\begin{pmatrix} I_x^2 & I_xI_y & I_xI_z \\ I_xI_y & I_y^2 & I_yI_z \\ I_xI_z & I_yI_z & I_z^2 \end{pmatrix}$$

- Local structure tensor (Weickert'98, Fernandez'03)



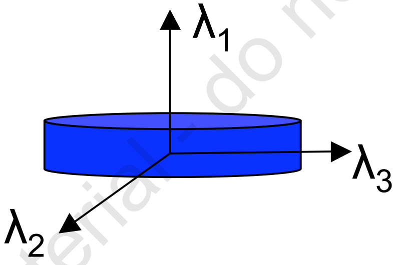
$$\begin{pmatrix} I_x^2 * G_\sigma & I_xI_y * G_\sigma & I_xI_z * G_\sigma \\ I_xI_y * G_\sigma & I_y^2 * G_\sigma & I_yI_z * G_\sigma \\ I_xI_z * G_\sigma & I_yI_z * G_\sigma & I_z^2 * G_\sigma \end{pmatrix}$$

- Property of local structure tensor



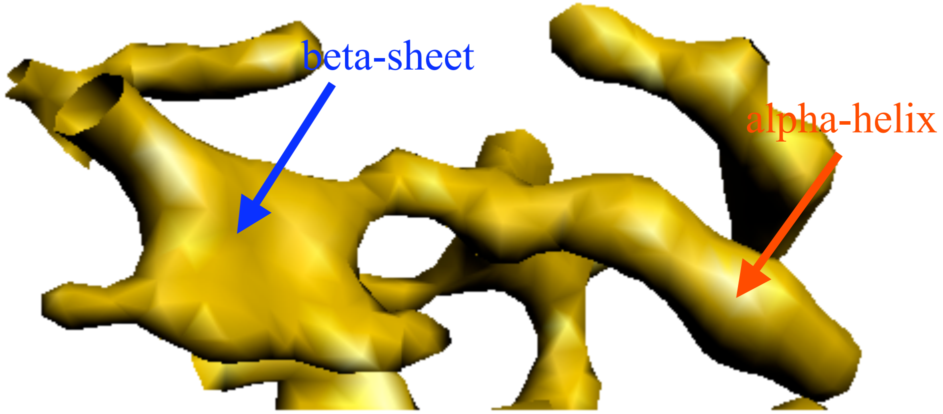
Line structure

$$\lambda_2 \approx \lambda_3 >> \lambda_1 \approx 0$$



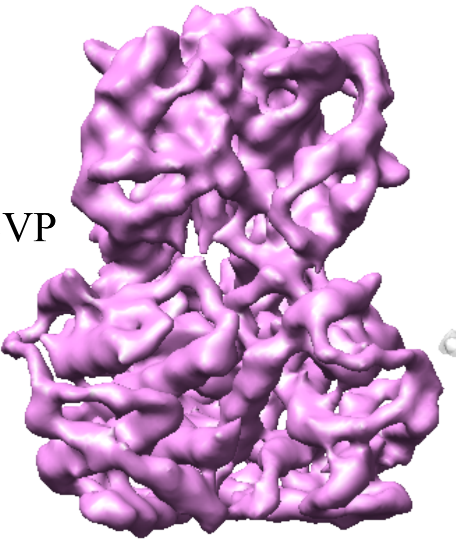
plane structure

$$\lambda_1 >> \lambda_2 \approx \lambda_3 \approx 0$$

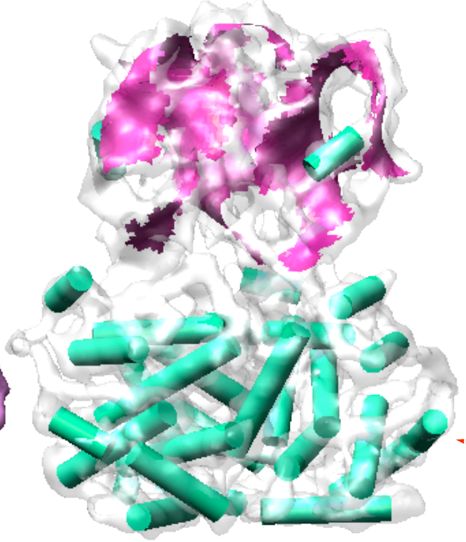


Step #5A: Quasi-Atomic Models

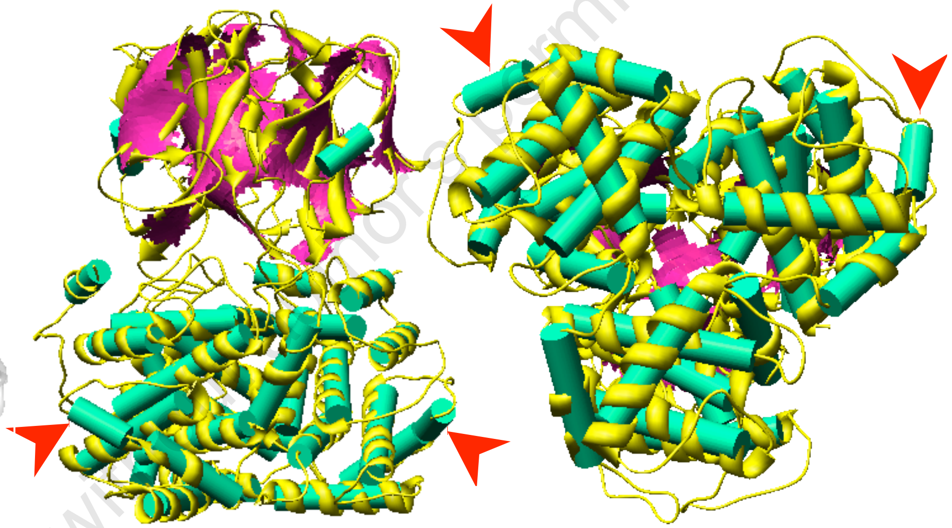
1BVP



Blurred map (8Å)

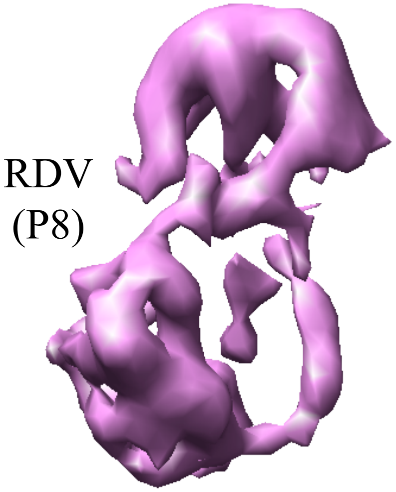


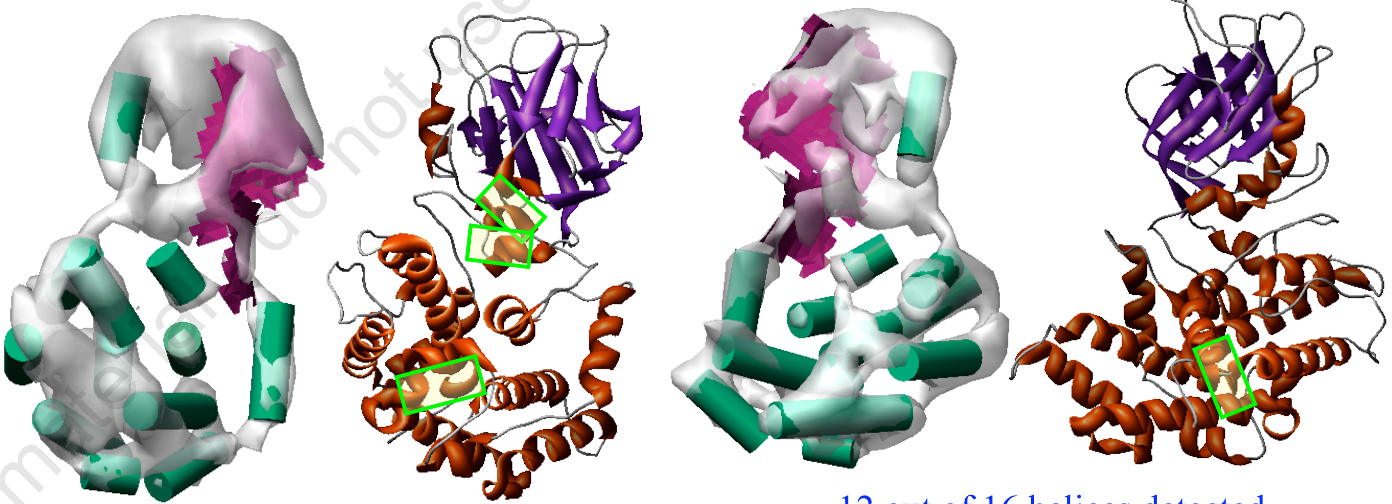
Helices-Sheets with map



Helices-Sheets with PDB (all 27 helices detected)

RDV
(P8)





12 out of 16 helices detected

Zeyun, Bajaj, IEEE/ACM Tran on CompBio&BioInf., 2007

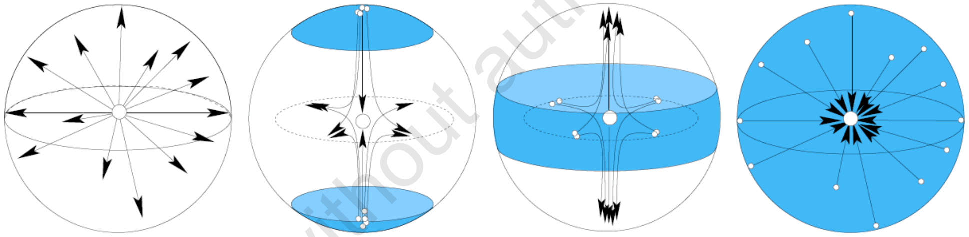
Critical Points, their Indices, and their Manifolds

Critical Point of a smooth function is a point where the gradient of the function vanishes.

Index of a critical point is the number of independent directions in which the function decreases.

In 3D, four types of critical points

1. Minima – index 0
2. Saddle of index 1
3. Saddle of index 2
4. Maxima – index 3



From EHNP, SoCG'03

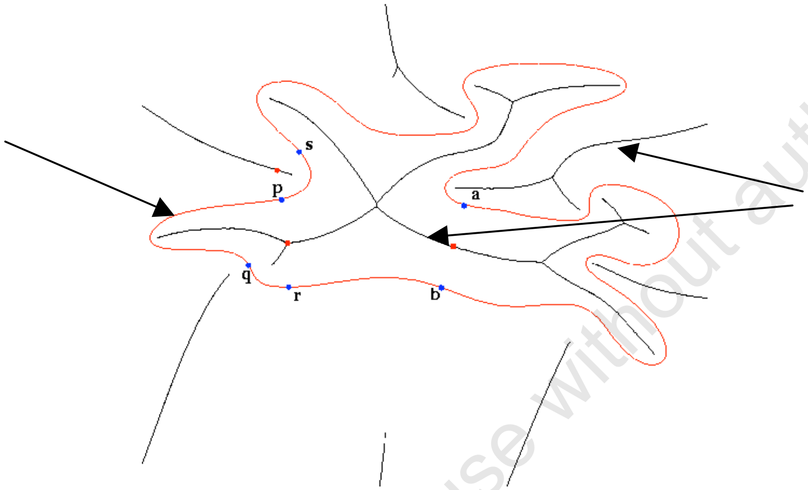
<u>Integral curve</u> : A path in the domain of the function on which at every point the tangent to the curve equals the gradient of the function.

<u>Stable Manifold</u> of a critical point is the union of all integral curves ending at the critical point.

<u>Unstable Manifold</u> of a critical point is the union of all integral curves starting at the critical point.

Medial Axis and Distance Functions

Medial Axis *M* of a shape *S* is defined as a set of points which has more than one nearest point on *S*.



Shape *S*

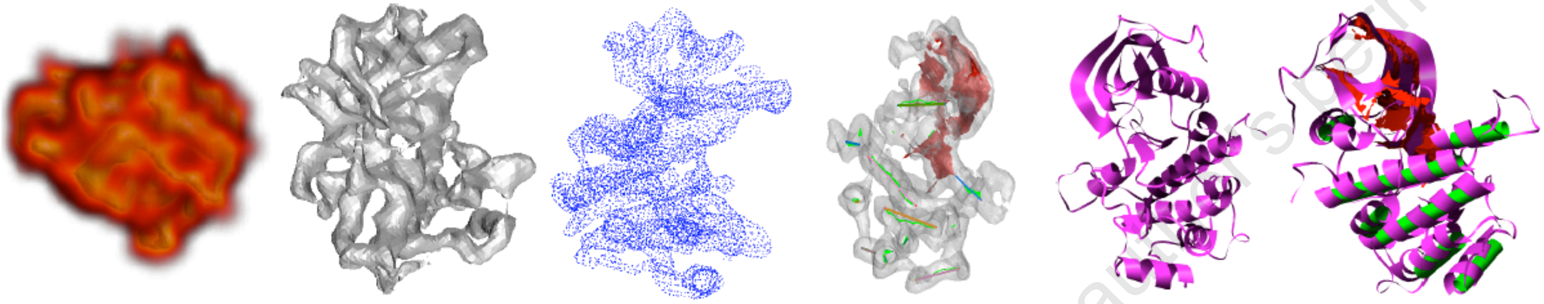
Medial Axis *M*

Distance Function h_S assigns every point *x* the nearest distance to *S*.

Approximation of h_S is done via h_P when *S* is known only via a finite set of points *P* on *S*.

$$h_P : \mathbb{R}^3 \to \mathbb{R}, \quad x \mapsto \min_{p \in P} \|x - p\|$$

#5B-I: Secondary Structure Elucidation



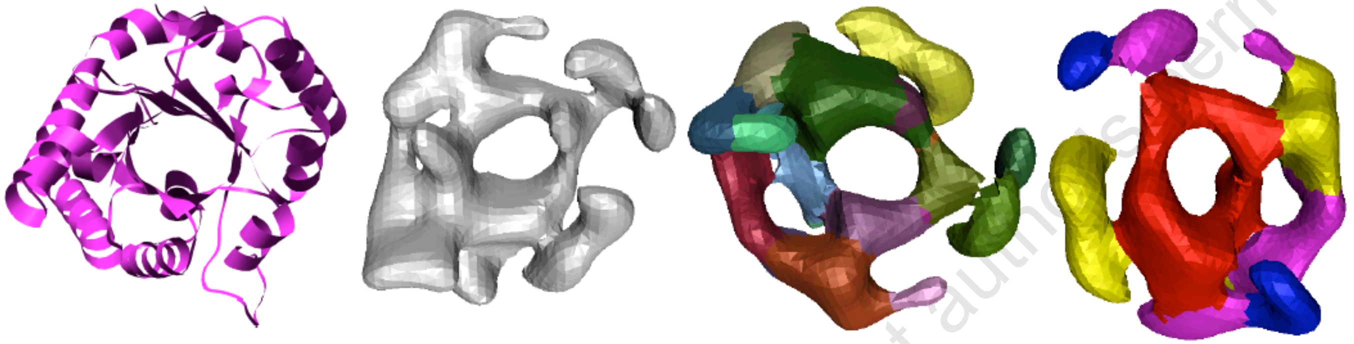
Step 1: Vor/Del(P) computation

Step 2: Identification of Interior Medial Axis M.

Step 3:
  3.a: Identification of Critical points of distance function from Vor/Del(P).
  3.b: Selection of Critical points only on M.

- Step 4: Classification of Medial Axis via
  - 4.a: U_1 – Unstable Manifold of index 1 saddle point on M
  - 4.b: U_2 – Unstable Manifold of index 2 saddle point on M.
- Step 5: Width Test to select the subsets of U_1 (β-sheets) and U_2 (α-helices).

#5B-II:Tertiary Structure Elucidation



Step 1: Vor/Del(P) computation

Step 2: Identification of Interior Medial Axis M

Step 3:
  3.a: Identification of Critical points of h_P from Vor/Del(P).
  3.b: Selection of Critical points only on M.

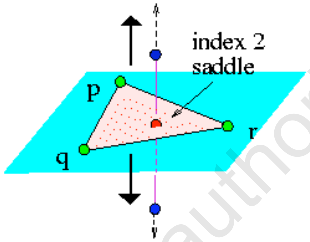
- Step 4: Decomposition of shape via S3 stable manifold of maxima on M
- Step 5: Width Test to select the subsets of S3 .

Computation of U_2

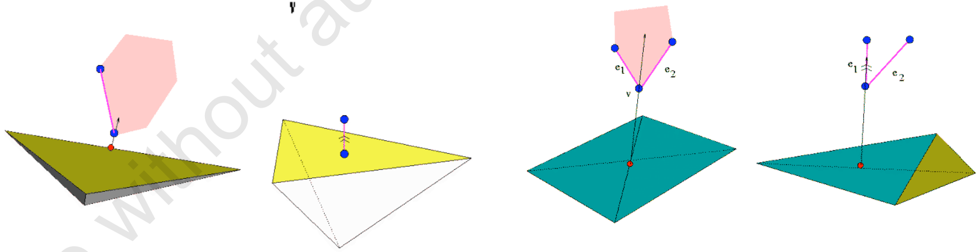
- Start from an index-2 saddle point.

- Generic step: Once reaching a VV, decide if the flow goes through a VE or if it enters a VF.

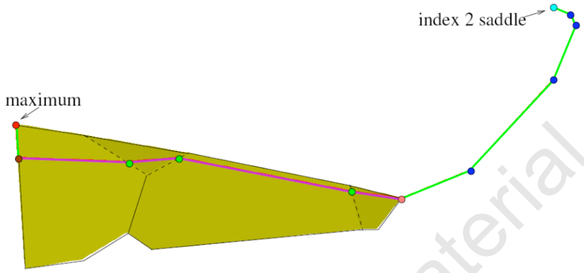
- Terminate at a local maximum.



Start



Generic Step



Structure of U_2

— Part of U_2 along Voronoi edge

— Part of U_2 through Voronoi facets

● Voronoi vertex on U_2

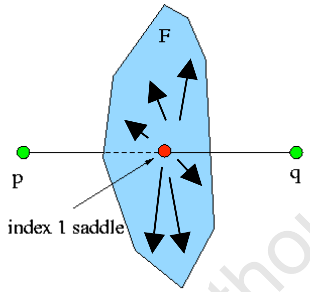
● Non-Voronoi vertex on U_2

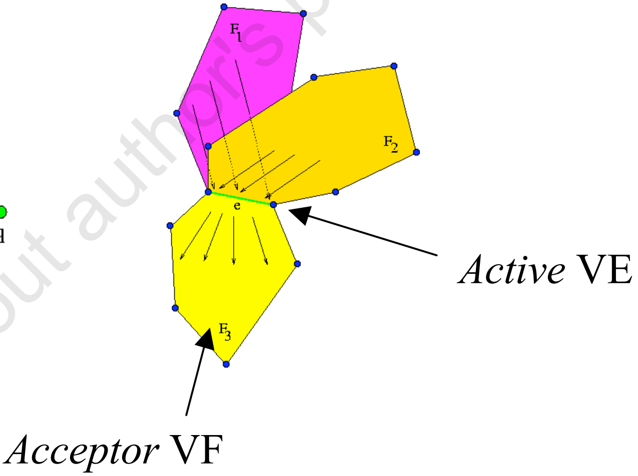
Goswami, Dey, Bajaj, ICVGIP'06

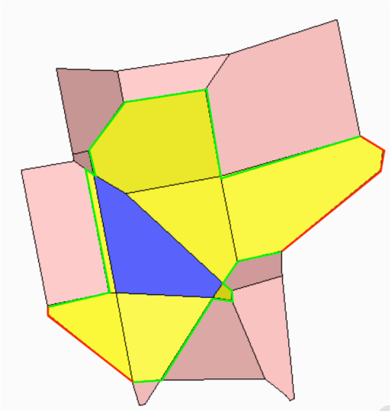
Computation of U_1

- Start from the VF containing an index-1 saddle point.

- Generic step: Identify the VE s on the boundary which are *active*. Expand the unstable manifold by collecting their *acceptor* VF s.





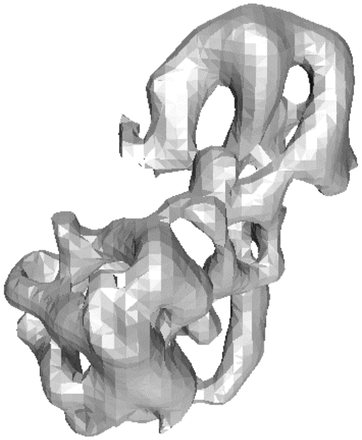


🟩 *Active* Voronoi edges to extend U_1

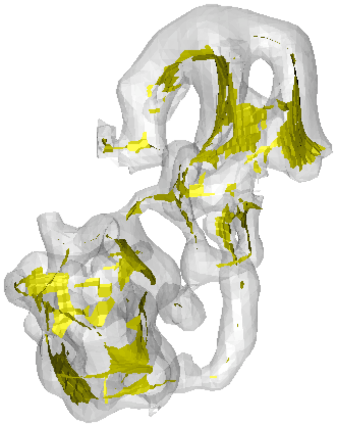
🟥 *Passive* Voronoi edges defining the boundary of U_1

Goswami, Dey, Bajaj, ICVGIP'06

Secondary Structure of RDV Outer Capsid Coat Protein P8

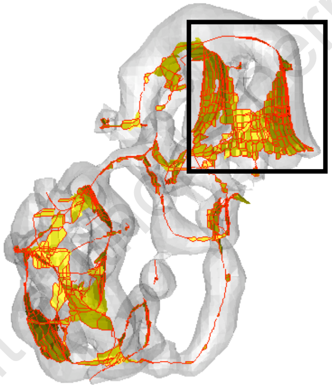


Surface

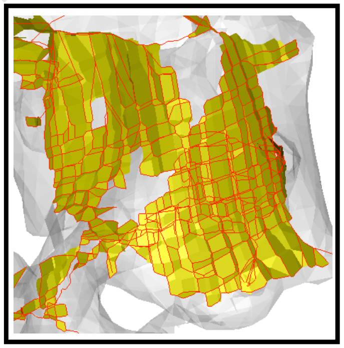


U_1

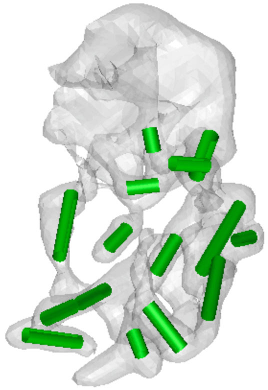
U_2



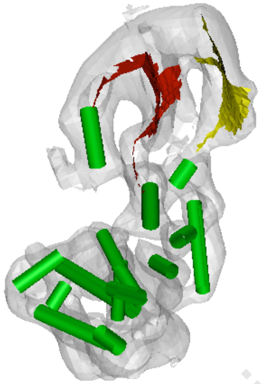
U_1 and U_2



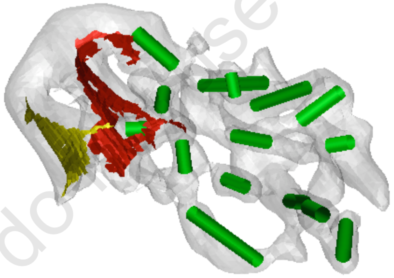
Closeup



Helices



Helices and Sheets



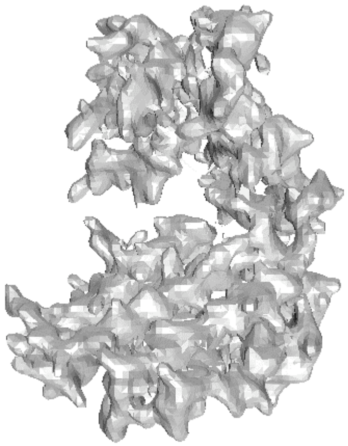
Another View

Ribbon Diag. of PDB

α-helices 15/16
β-sheets 3/3

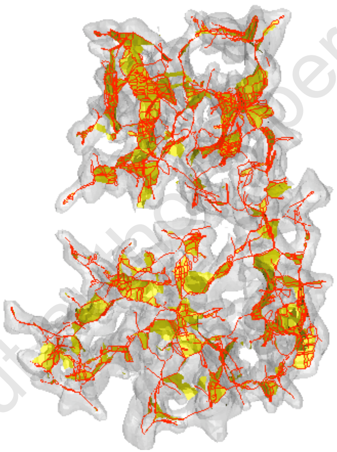
3D EM map of P8 segmented out from cryo-EM map of Rice Dwarf Virus (RDV) at 6.8 A resolution

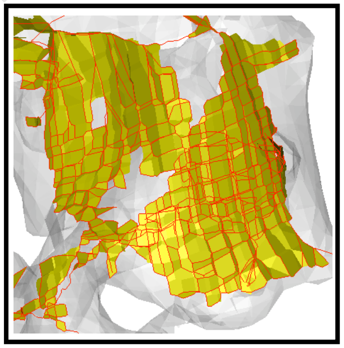
Secondary Structure of Bacterial Chaperonin GroEL

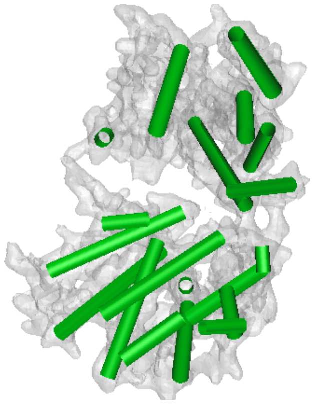

Surface

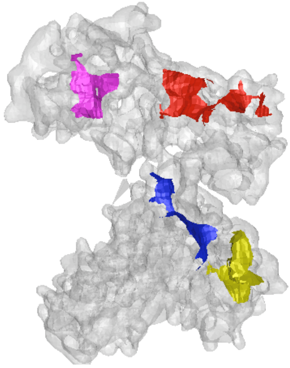

U$_1$

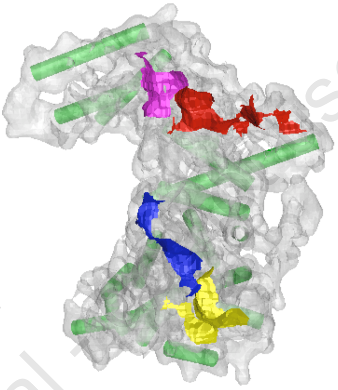

U$_2$


U$_1$ and U$_2$

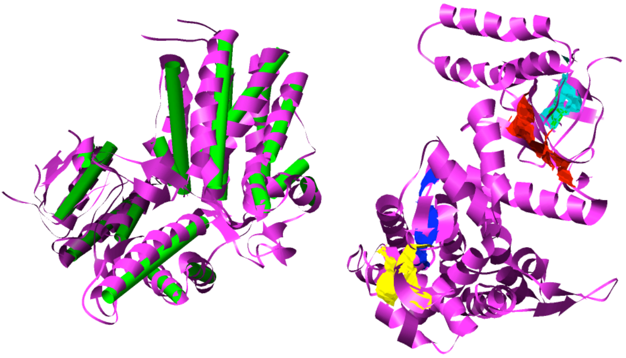

Closeup


Helices


Sheets


Helices and Sheets


Ribbon Diag. of PDB


Superimposed

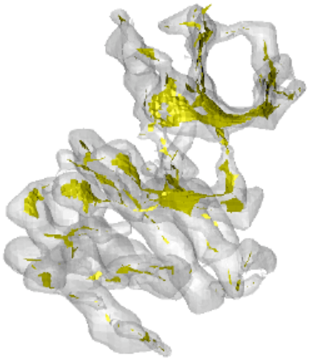
α-helices 15/16

β-sheets 3/3

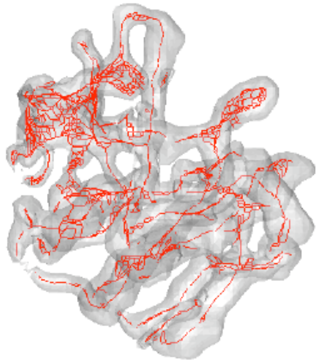
Center for Computational Visualization http://www.ices.utexas.edu/CCV
Institute of Computational and Engineering Sciences
Department of Computer Sciences
University of Texas at Austin

May 2007

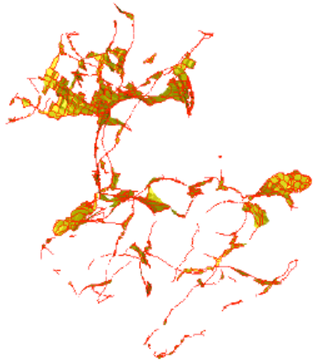
Width Test



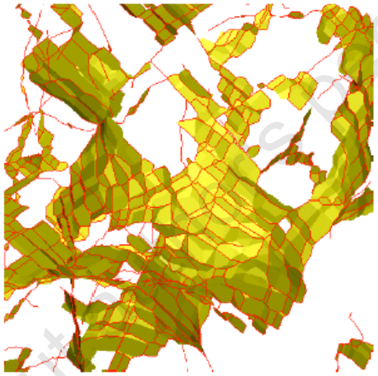
U_1



U_2



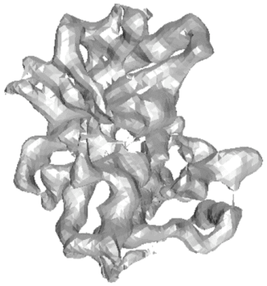
U_1 and U_2

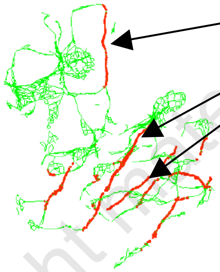


Closeup: U_1 is bounded by U_2

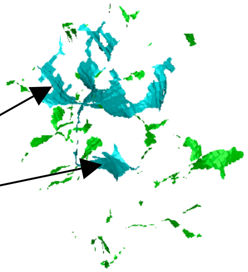
- U_1 and U_2 give superset of sheets and helices.
- α-helix: width 2.5 A and pitch-length 1.5 A [Branden-Tooze]
- β-sheet: thickness 1.5 A [Branden-Tooze]

- h_P values of Voronoi elements constituting U_1 and U_2 help select the subset that passes the width and thickness test.



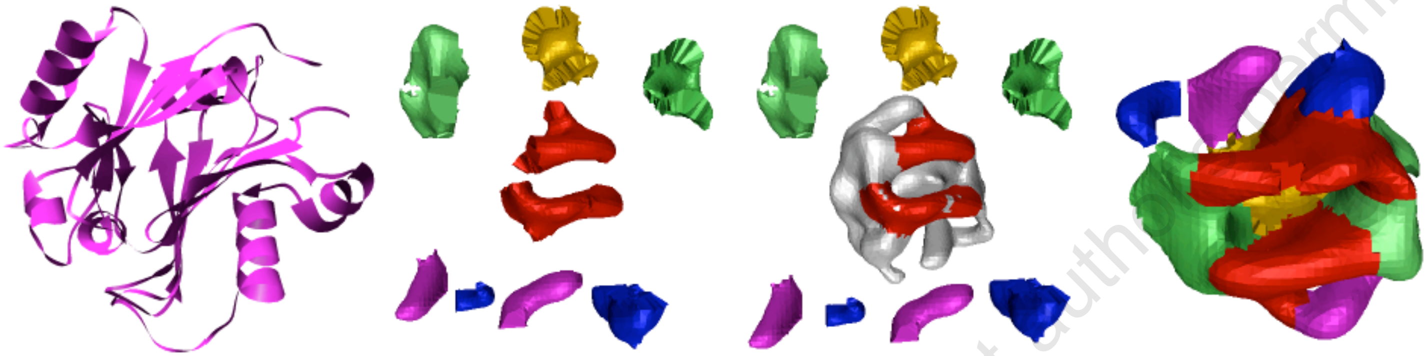


Subset of U_2 that corresponds to helices



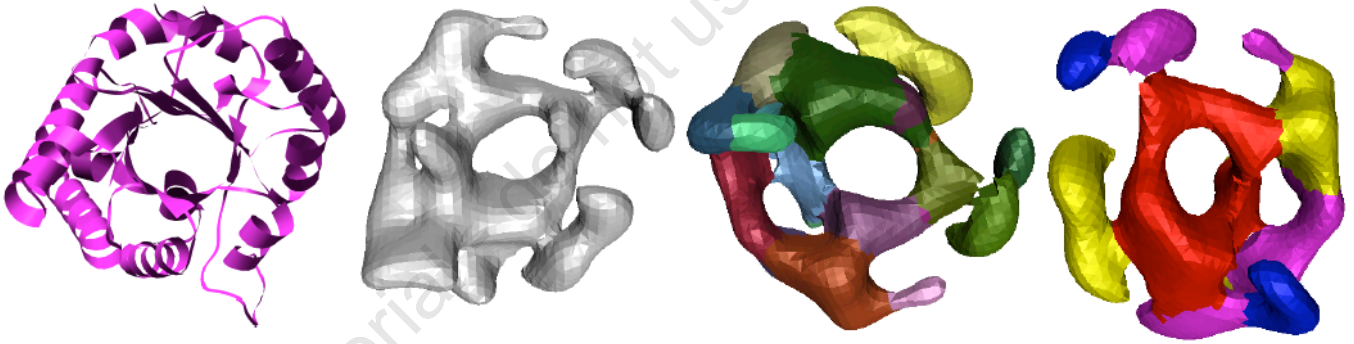
Subset of U_1 that corresponds to sheets

Tertiary Structure Elucidation


(a) (b) (c) (d)

1AOR (12Å)

The tertiary fold of 1AOR is a β-sandwich (two red sheets), which is surrounded by the differently colored helical segments.



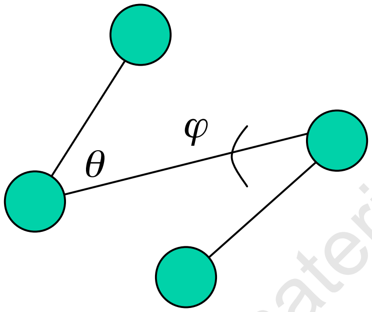
The tertiary fold of 1TIM is .a α/β-barrel.
The β -region in the middle is segmented as red while the helical segments surrounding it are colored differently.

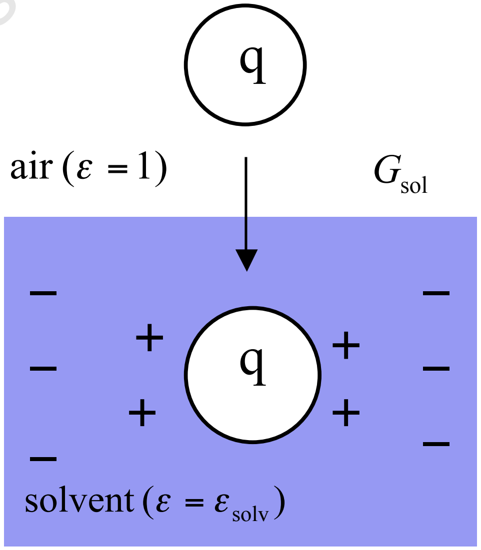
Free Energy of a Macromolecule in Solvent
(Towards Flexible Models of Proteins)

Total free energy : $G = \boxed{E_{MM}} + \boxed{G_{\text{sol}}} - TS$

$$E = E_b + E_\theta + E_\varphi + E_{vdw} + E_{elec}$$

$$G_{\text{sol}} = G_{\text{cav}} + G_{\text{vdw}} + \boxed{G_{\text{pol}}}$$





How do we Compute G_{pol} ?

$$G_{\text{pol}} = \frac{1}{2}\int [\phi_{\text{solvent}}(\mathbf{r}) - \phi_{\text{air}}(\mathbf{r})]\rho(\mathbf{r})\,dV$$

ε	dielectric coefficients
ϕ	electrostatic potential
ρ	solute charge density
λ	ion accessibility parameter
c_j^{∞}	ion bulk concentration
q_j	ion charge
k_B	Boltzmann's constant
T	temperature

- (a) Poisson-Boltzmann (PB) Theory

$$-\nabla\cdot[\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = 4\pi\rho(\mathbf{r}) + 4\pi\lambda(\mathbf{r})\sum_{j=1} c_j^{\infty} q_j \exp(-q_j\phi(\mathbf{r})/k_B T)$$

Molecular volume V for charge density, dielectric interface

- (b) Generalized Born (GB) Theory
  - Born formula (Born 1920), Generalized Born formula (Still 1990)

$$G_{\text{pol}} = -\frac{\tau}{2}\sum_{ij}\frac{q_i q_j}{[r_{ij}^2 + R_i R_j \exp(-\frac{r_{ij}^2}{4R_i R_j})]^{\frac{1}{2}}}$$

$$R_i^{-1} = \frac{1}{4\pi}\int_{\text{ex}} \frac{1}{|\mathbf{r} - \mathbf{x}_i|^4}\,dV$$

$\tau = 1 - \frac{1}{\varepsilon_{\text{solv}}}$ r_{ij} : distance between atom i and j

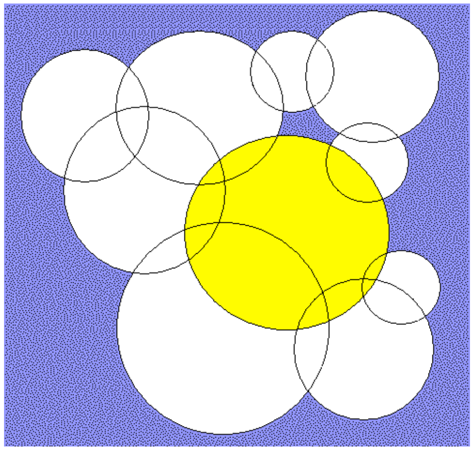
ex: exterior to molecular surface

q_i : charge of atom i R_i : effective Born radius of atom i

Step 6b: GB based G_{pol} - Calculation of Born Radii

A single charge q_i located at the center of atom i in the molecule.

$$G_{\text{pol}} = -\frac{\tau}{2}\frac{q_i^2}{R_i} \qquad \text{(by GB)}$$



On the other hand, by approximating the electric field as the Columbic field,

$$G_{\text{pol}} = -\tfrac{1}{8\pi}\tau\int_{\text{ex}} \frac{q_i^{\,2}}{|\mathbf{r}-\mathbf{x}_i|^4}\,dV$$

Therefore $$R_i^{-1} = \frac{1}{4\pi}\int_{\text{ex}} \frac{1}{|\mathbf{r}-\mathbf{x}_i|^4}\,dV$$

ex: exterior of the molecule

via Gauss' Divergence Thm

$$R_i^{-1} = \frac{1}{4\pi}\int_{\Gamma} \frac{(\mathbf{r}-\mathbf{x}_i)\cdot\mathbf{n}(\mathbf{r})}{|\mathbf{r}-\mathbf{x}_i|^4}\,dS$$

Γ : molecular surface

Step 6b: GB based G_{pol} - Fast Calculation of Born Radii

$$R_i^{-1} = \frac{1}{4\pi}\int_\Gamma \frac{(\mathbf{r}-\mathbf{x}_i)\cdot\mathbf{n}(\mathbf{r})}{|\mathbf{r}-\mathbf{x}_i|^4}\,dS \approx \frac{1}{4\pi}\sum_{k=1}^{N} w_k \frac{(\mathbf{r}_k-\mathbf{x}_i)\cdot\mathbf{n}(\mathbf{r}_k)}{|\mathbf{r}_k-\mathbf{x}_i|^4}, \qquad \mathbf{r}_k \in \Gamma$$

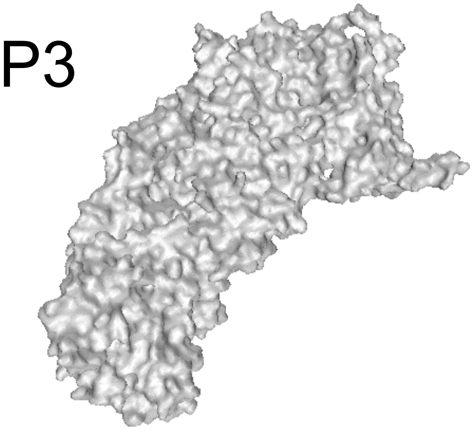
Algorithm:

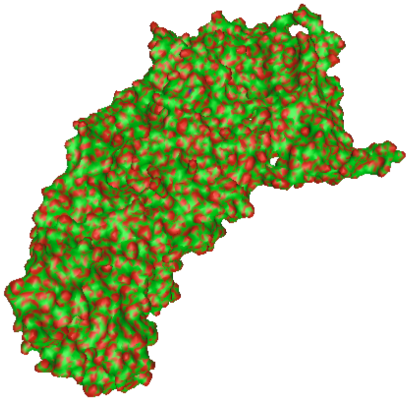
1. Generate a model for the molecular surface Γ.
2. Cubature: choose w_k and $\mathbf{r}_k$ properly so that higher order accuracy can be obtained for small N.
3. Fast Fourier summation to evaluate R_i $i = 1,\ldots,M$.

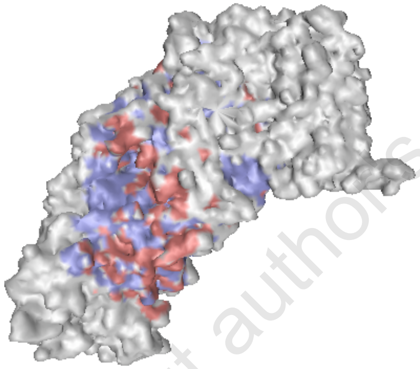
Bajaj, Zhao 2007

Shape/Energetic Properties of RDV capsid proteins

P3

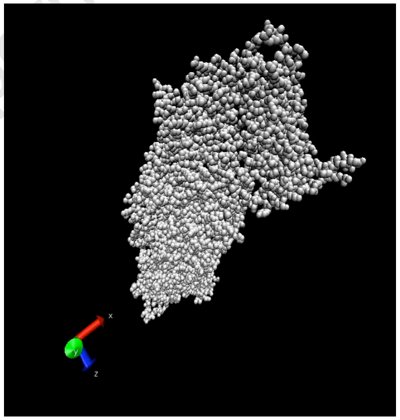

Molecular Surface


Gaussian Curvature

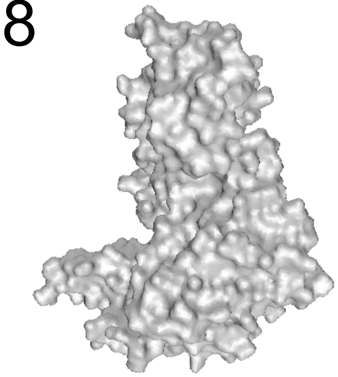

spline elements 75416

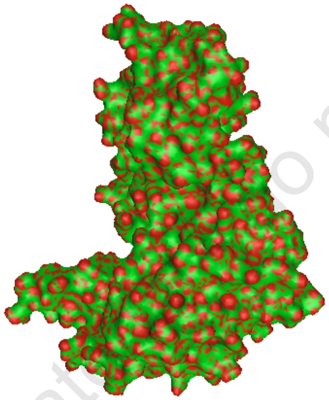
Slide 0

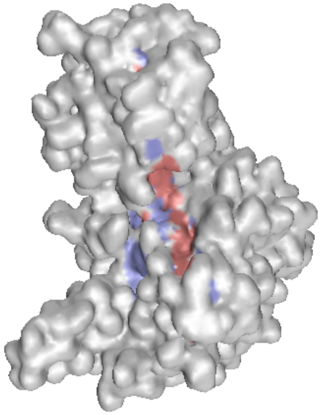
GB Gpol:
-23892.33 kcal/Mol

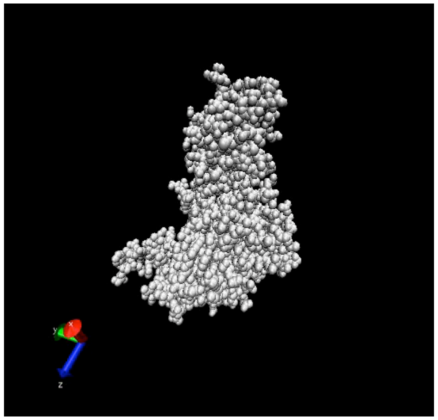


P8


Molecular Surface


Gaussian Curvature

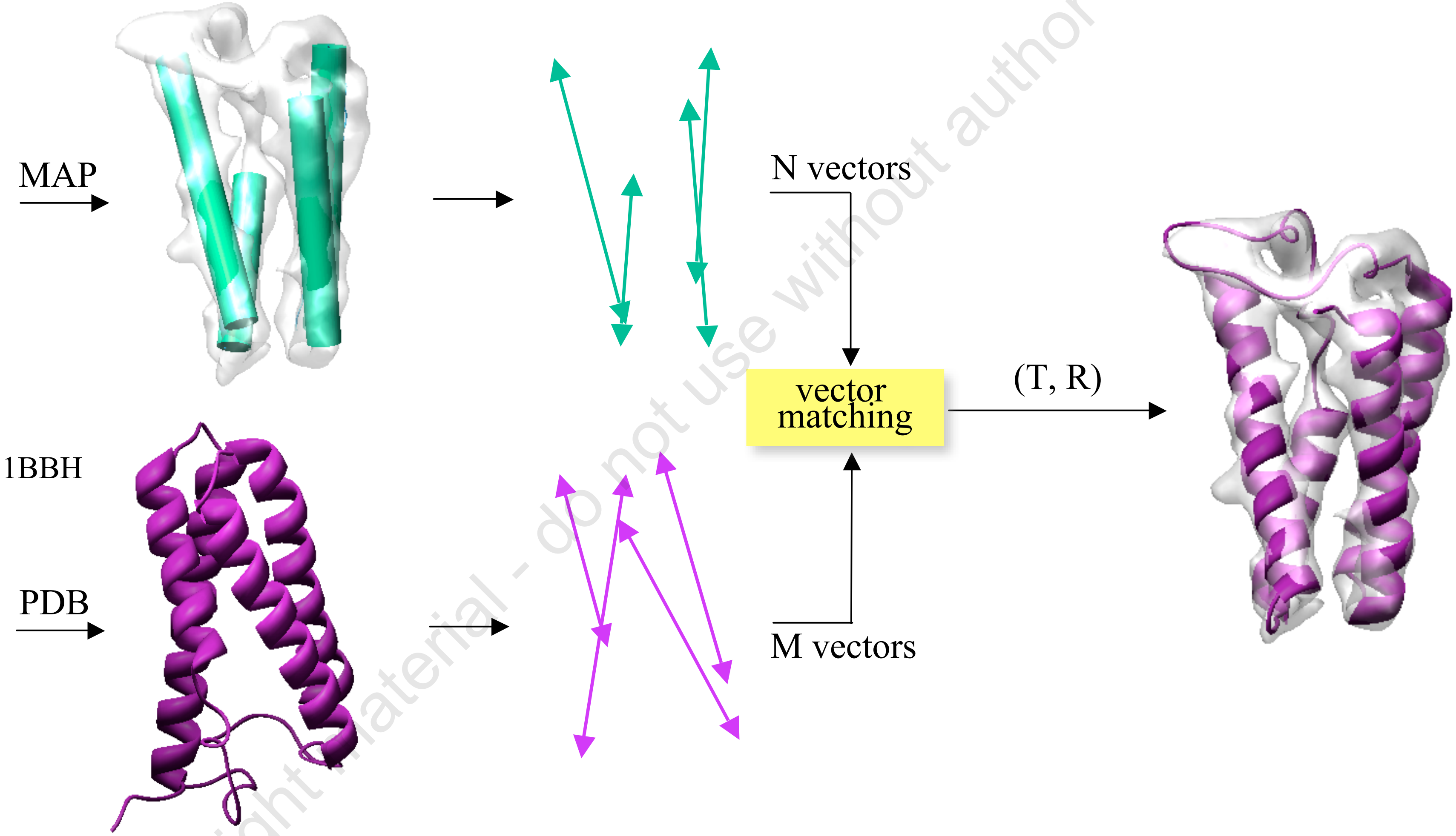

spline elements 66088
GB Gpol:
-8413.185 kcal/Mol



Zhao, Xu, Bajaj, ACM/IEEE SPM'07

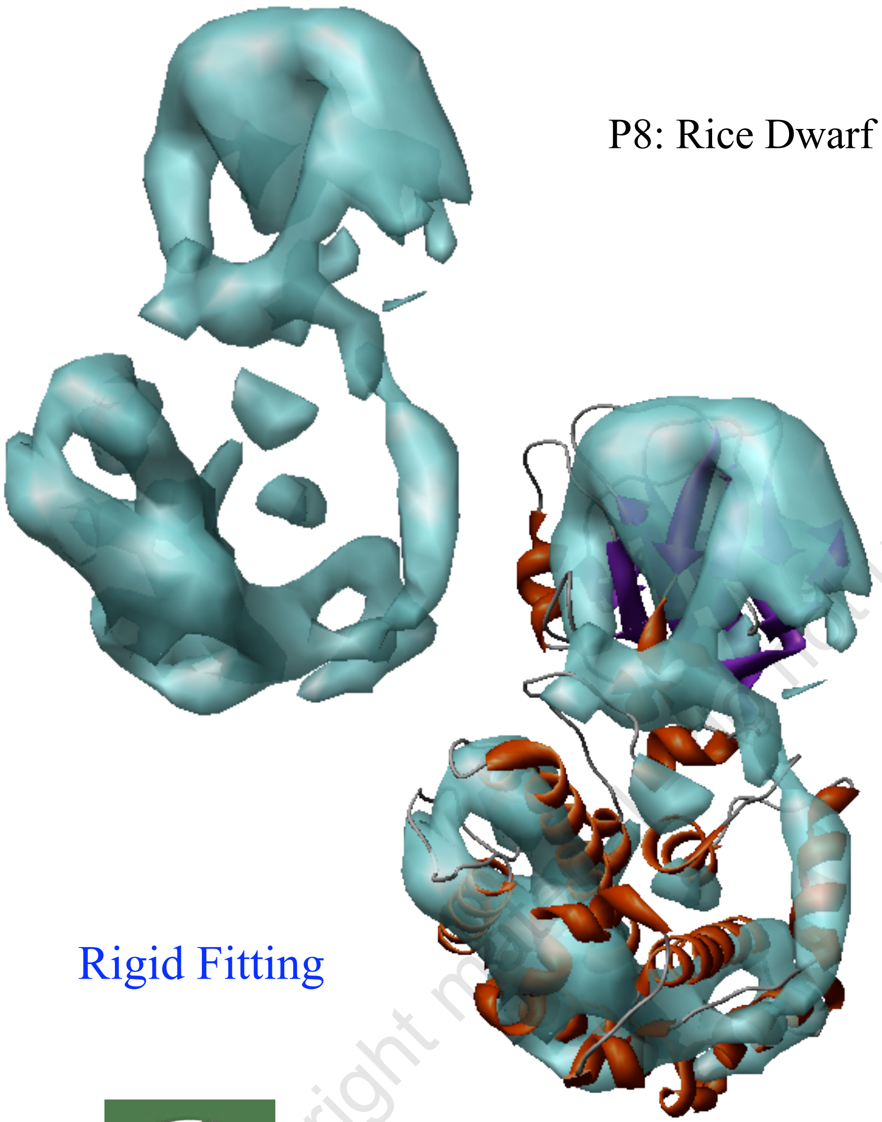
Step 7a: Match and Fit using elucidated structural elements

- vector matching approach

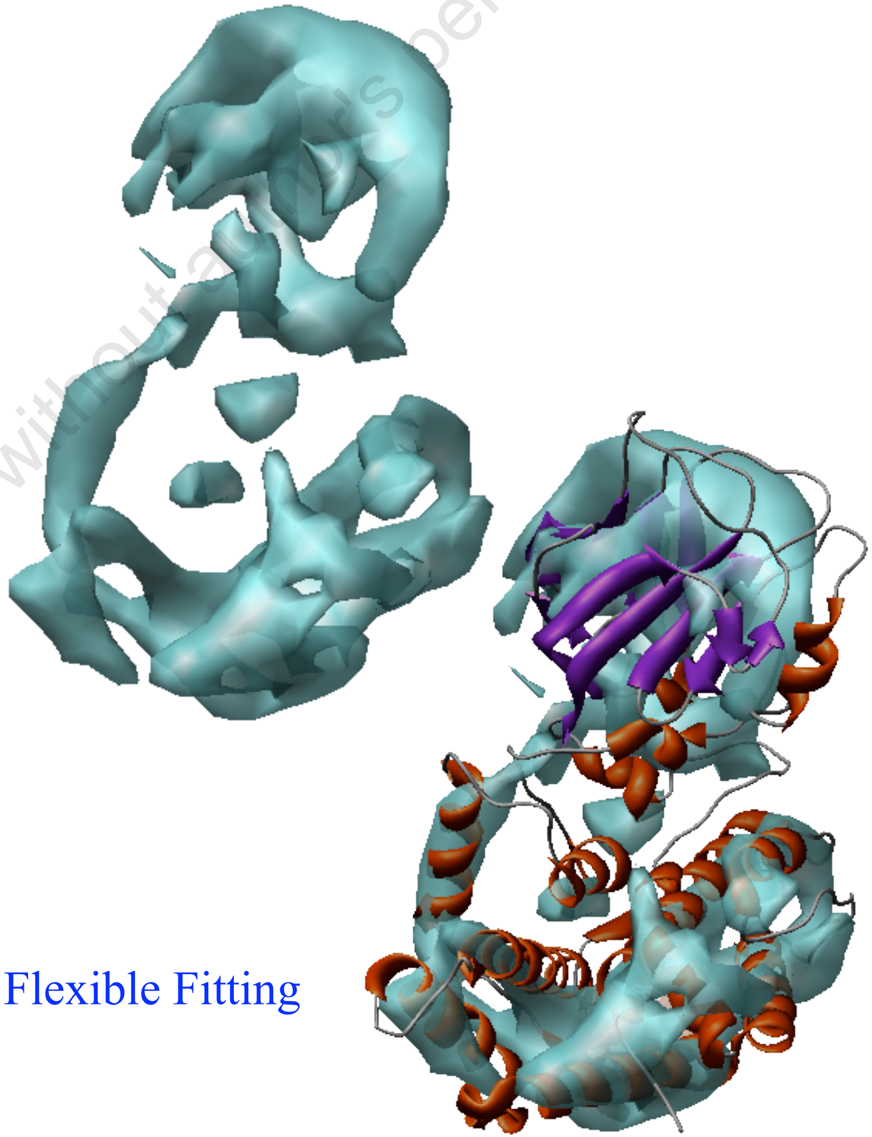


Step 7b: Flexible Match and Fit using generalized modal analysis

P8: Rice Dwarf Virus



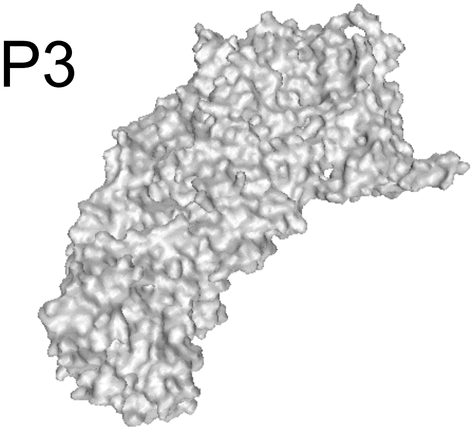
Rigid Fitting



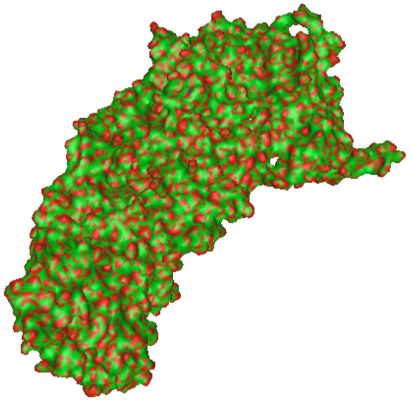
Flexible Fitting

Shape/Energetic Properties of RDV capsid proteins

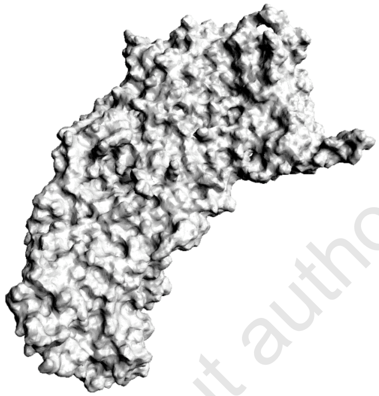
P3



Molecular Surface

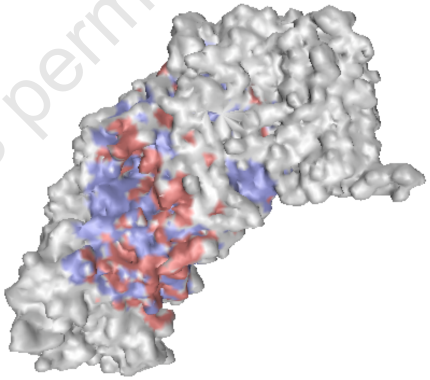


Gaussian Curvature



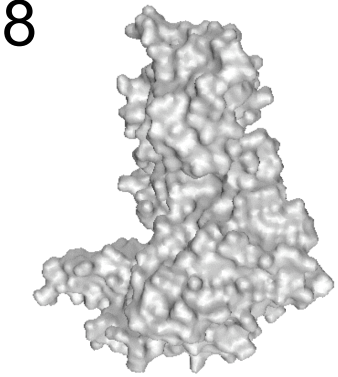
spline elements 75416

GB Gpol:
-23892.33 kcal/Mol

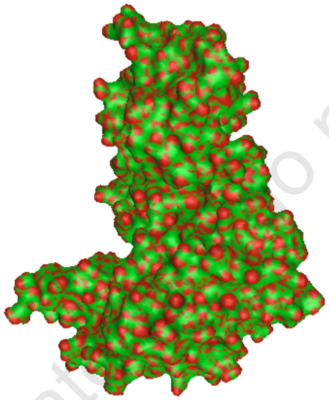


PB electrostatics:

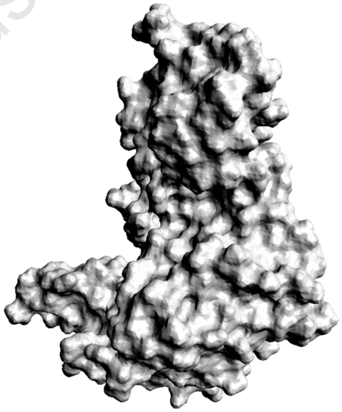
P8



Molecular Surface

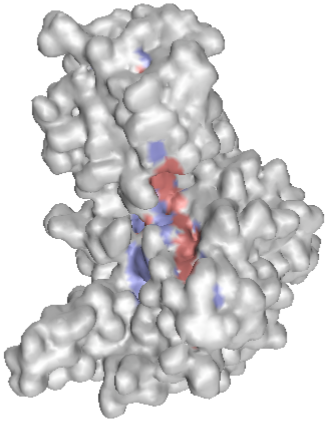


Gaussian Curvature



spline elements 66088

GB Gpol:
-8413.185 kcal/Mol

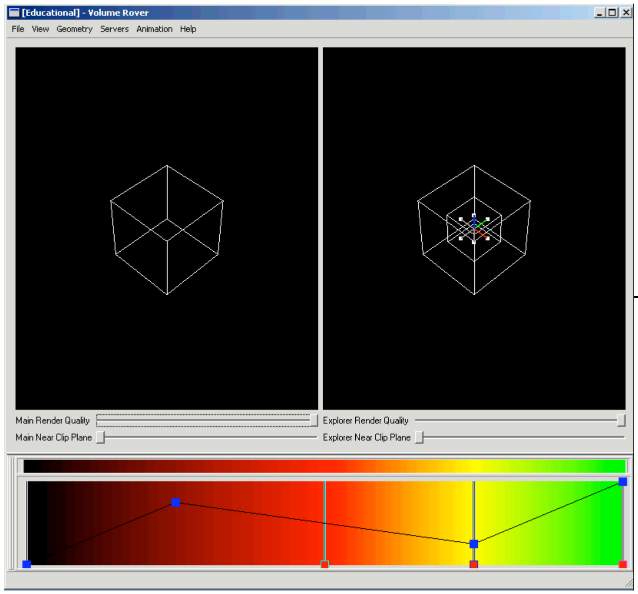


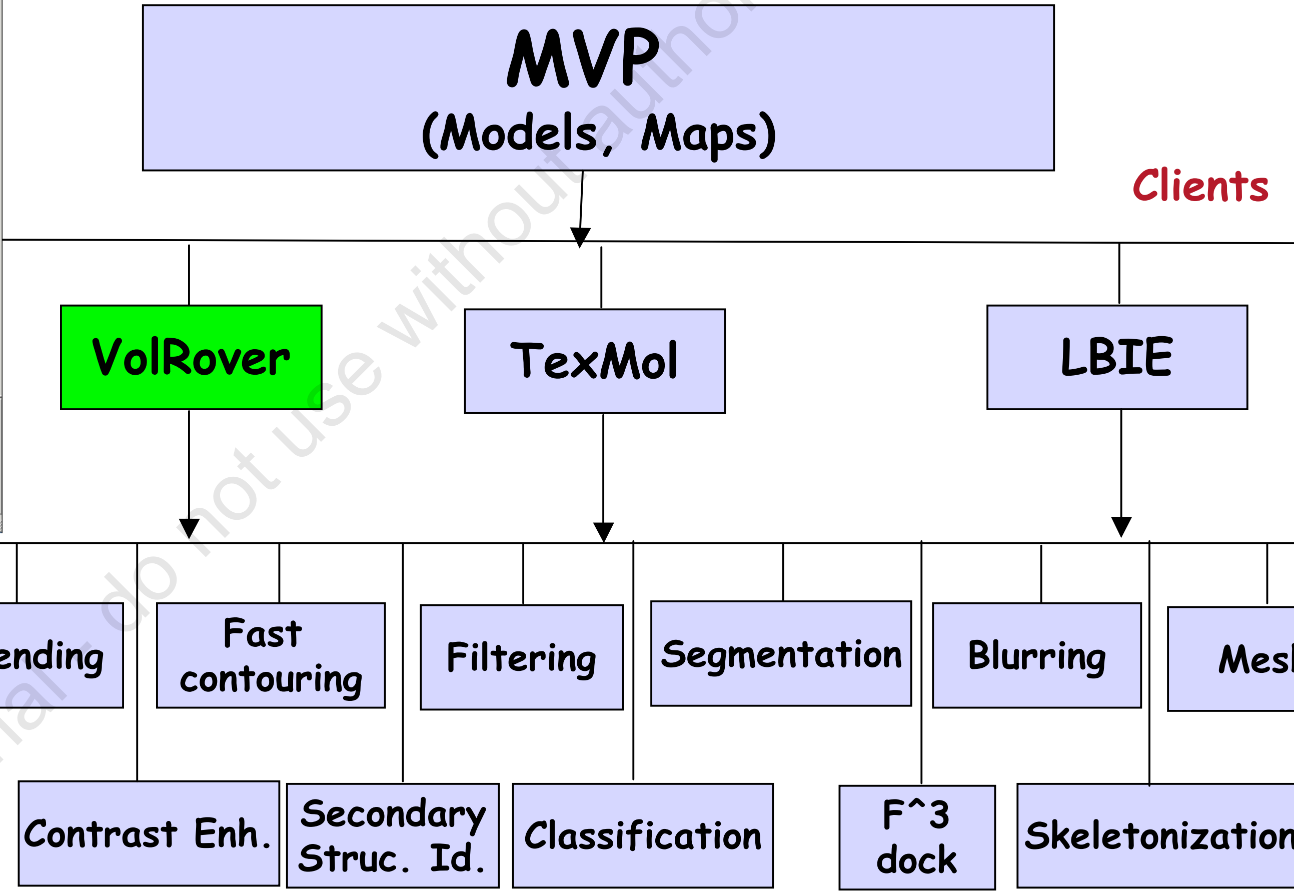
PB electrostatics:

CCV Software

Open Source and in Public Domain

(http://www.ices.utexas.edu/CCV/software/)





MVP
(Models, Maps)

Clients

VolRover

TexMol

LBIE

VolRending

Fast contouring

Filtering

Segmentation

Blurring

Mesl

Libraries

(server)

Contrast Enh.

Secondary Struc. Id.

Classification

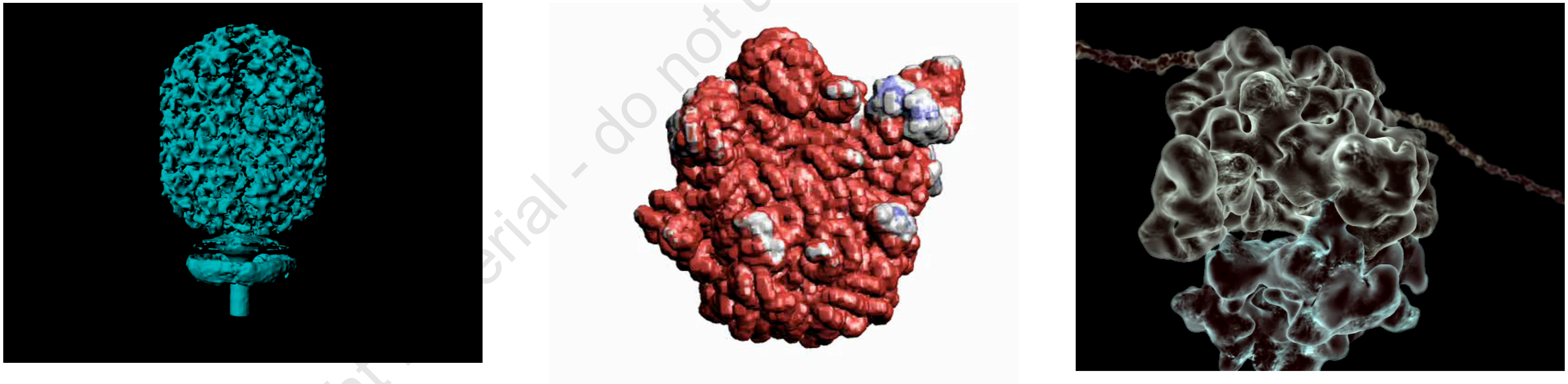
F^3 dock

Skeletonizatio

Acknowledgements

- ***Group Members***
  - **Albert Chen (CS, Ph.D)**
  - **Andrew Gillete (Math, Ph.D.)**
  - **Samrat Goswami (PostDoc)**
  - **Zeyun Yu (UCSD)****
  - **Wenqi Zhao (CAM, Ph.D.)**

- ***Senior Collaborators***
  - Tim Baker (UCSD)
  - Nathan Baker (WashU)
  - Wah Chiu (Baylor)
  - Andy McCammon (UCSD)
  - Art Olson (Scripps)
- ***Sponsored by***
  - **NIH: P20-RR020647, R01-GM074258, R01-EB004873**
  - **NSF: ITR-IIS-032550**



Acknowledgements

- ***Group Members***
  - Albert Chen (CS, Ph.D)
  - **Katherine Clarridge (MBE,MD)****
  - **Tamal Dey (OSU) ****
  - Andrew Gillete (Math, Ph.D.)
  - Samrat Goswami (PostDoc)
  - **Inwoo Ha (CS, MS)**
  - **Insung Ihm (SU,S. Korea)****
  - **Sangmin Park (CS, Ph.D.)**
  - **Myung-Soo Kim (SNU,S. Korea)****
  - **Bong-June Kwon (CS, M.S.)**
  - **Bong-Soo Sohn (KBU,S. Korea)****
  - **Jason Sun (Res. Scientist)**
  - **Vinay Siddahanavalli (CS, Ph.D.)**
  - **Jesse Sweet (Math, Ph.D.)**
  - **Mason Weems (CMB, Ph.D.)**
  - **Guoliang Xu (AS, China)****
  - Zeyun Yu (UCSD)**
  - **Xiaoyu Zhang (CSU)****
  - **Jessica Zhang (PostDoc)**
  - Wenqi Zhao (ICES, Ph.D.)
- ***Senior Collaborators***
  - Manfred Auer (LBL)
  - Nathan Baker (Wash. U.)
  - Tim Baker (UCSD)
  - Tom Bartol (Salk)
  - Wah Chiu (Baylor)
  - Gregory Gladish, J. Hazle (MD Anderson)
  - Kirsten Harris (Neuro, UT)
  - Tom Hughes (ICES)
  - Steve Ludtke (Baylor)
  - Andy McCammon (USCD)
  - Tinsley Oden (ICES)
  - Alberto Paoluzzi (U of Roma, Tre)
  - Art Olson, M. Sanner (Scripps)
  - Peter Rossky (ICES)
  - Andre Sali (UCSF)
- ***Funding***
  - **NIH: P20-RR020647-01, R01-GM074258, R01-GM073087, R01-EB004873**
  - **NSF: ITR-IIS-032550, DDDAS-CNS-0540033**
  - **UT: Ti-3D**