A few Applications of Pattern Recognition Techniques to Proteomics

VIRGINIO CANTONI

Dep. Industrial and Information Engineering

Università di PAVIA

virginio.cantoni@unipv.it

Overview

- Searching in a database of protein structures
    - Search for a structural "motif"
    - Pairwise comparison
    - All-to-All comparison
- Study the interaction between structures and other molecules (Protein Docking)
    - Molecular surface representation
    - Protein surface comparison
    - Geometric shape descriptors
    - Shape matching algorithms

PDB

Protein Data Bank (PDB)
http://www.rcsb.org/pdb/

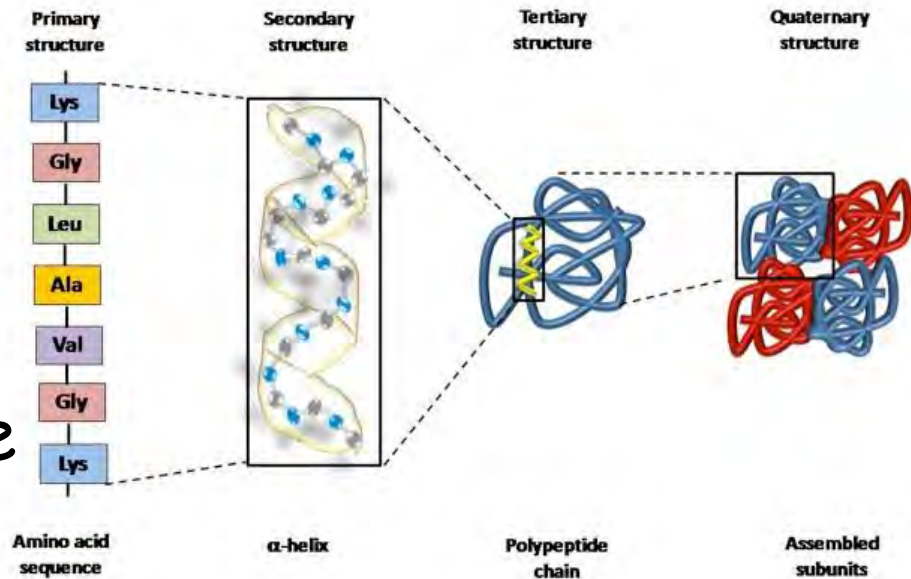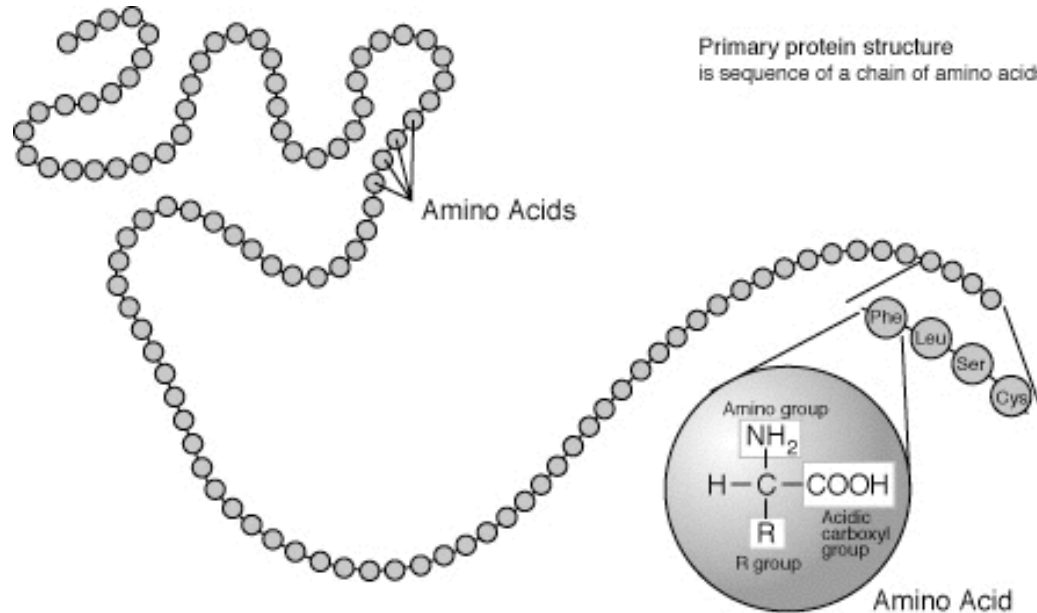| ATOM | 1 | N | GLY A | 1 | 44.842 | 51.034 | 101.284 | 0.01 | 27.20 |
ATOM	2	CA	GLY A	1	45.640	50.230	100.389	0.01	26.99
ATOM	3	C	GLY A	1	46.692	49.648	101.308	0.01	26.80
ATOM	4	O	GLY A	1	46.895	50.222	102.381	0.01	26.91
ATOM	5	N	SER A	2	47.283	48.516	100.951	1.00	26.26
ATOM	6	CA	SER A	2	48.277	47.866	101.761	1.00	26.17
ATOM	7	C	SER A	2	49.212	47.031	100.845	1.00	24.21
ATOM	8	O	SER A	2	49.060	47.195	99.630	1.00	19.77
ATOM	9	CB	SER A	2	47.438	47.091	102.800	1.00	26.31
ATOM	10	OG	SER A	2	46.276	46.356	102.404	1.00	27.99
ATOM	11	N	HIS A	3	50.147	46.186	101.370	1.00	23.93
ATOM	12	CA	HIS A	3	51.129	45.389	100.609	1.00	21.44
ATOM	13	C	HIS A	3	50.953	43.905	100.849	1.00	20.32
ATOM	14	O	HIS A	3	50.530	43.595	101.950	1.00	22.00
ATOM	15	CB	HIS A	3	52.555	45.674	100.990	1.00	19.69
ATOM	16	CG	HIS A	3	52.940	47.090	100.611	1.00	21.44
ATOM	17	ND1	HIS A	3	53.371	47.470	99.422	1.00	20.87
ATOM	18	CD2	HIS A	3	52.956	48.175	101.433	1.00	21.69
ATOM	19	CE1	HIS A	3	53.676	48.730	99.476	1.00	20.57

Levels of protein structure representation

- Primary structure
- Secondary structure
- Tertiary structure
- Quaternary structure

Primary structure: the sequence of amino acids

Primary protein structure is sequence of a chain of amino acids

Amino Acids

Phe Leu Ser Cys

Amino group
NH_2

$H-C-COOH$ — Acidic carboxyl group
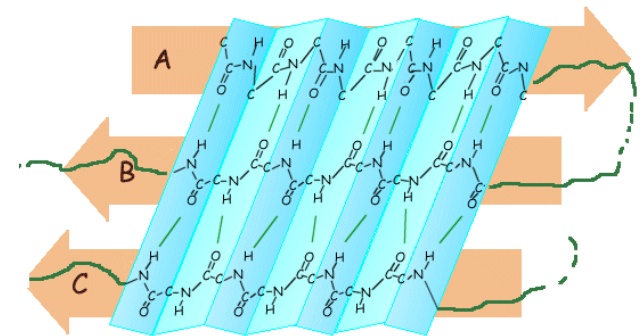
R
R group

Amino Acid

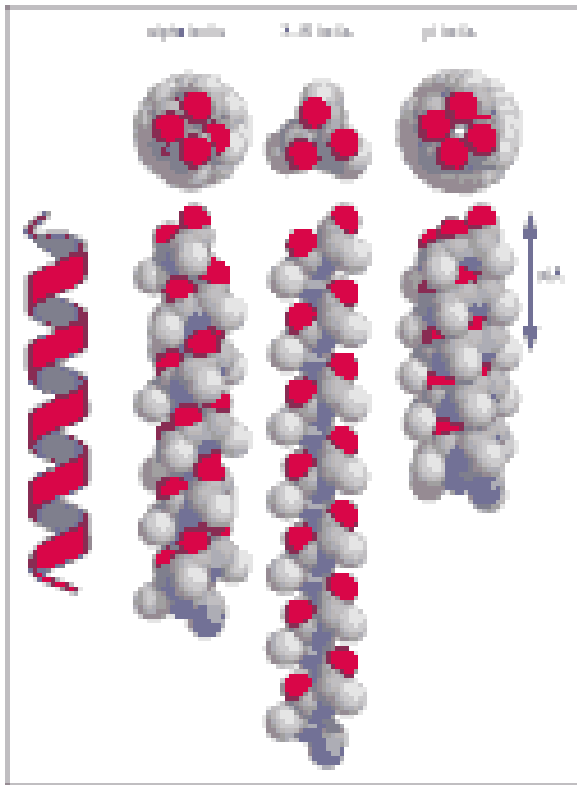MHGAYRTPRSKTDAYGCQILETRAS

Secondary structures

Three basic components:

- helix

- sheet

- Loops (linear connections between the components)
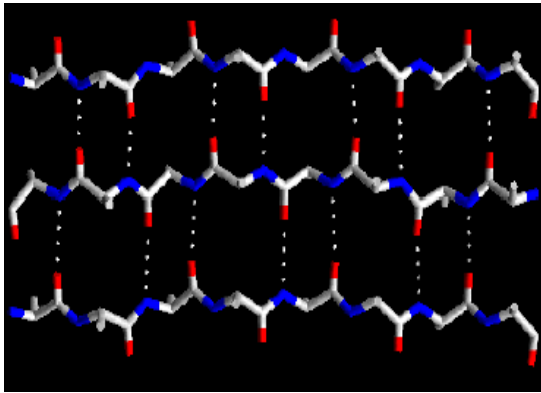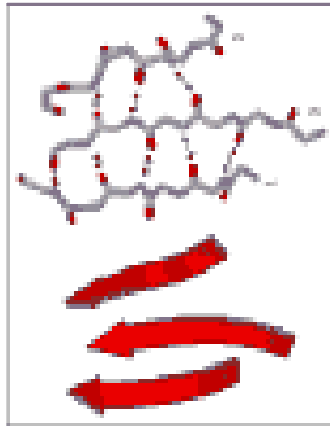
The helices

- One of the most closely packed arrangement of residues.
- ~40% of residues in globular proteins
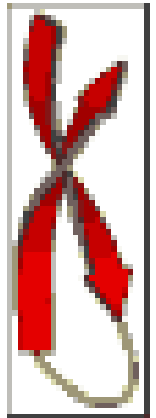
The sheet

loosely packed arrangement of residues.


Parallel

Antiparallel

Twisted

Secondary Structures Representation

- Secondary structures are represented as linear vectors (segments):

  the axis for the helix and the best fit segment for a sheet

- An alignment algorithm is used to match an helix segments with known axes to determine helix axis. Direct segment fits are made to fit sheet strands.

Secondary Structure Determination

- Programs: DSSP and STRIDE.

- On the average 4.8% of the target residues were differently assigned, this number reaching 12% for certain targets.

Protein Structure Comparison
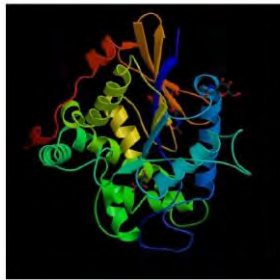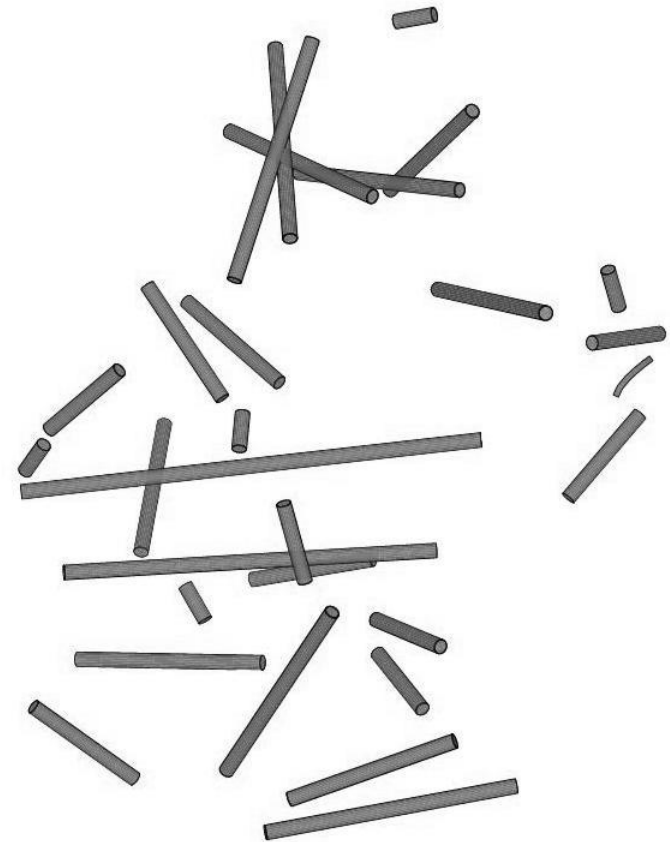

New protein

What are the most similar folds ?

PDB

Secondary structure representation

- Each secondary structure is displayed as a cylinder
- The protein is represented by and ordered sequence of cylinder with two labels: helices or sheets
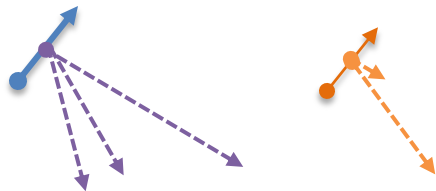
GHT applied to proteins

- For every protein, the distance (ρ) of every secondary structure from a reference point (RP, eg the geometric center of the protein) and the angle (theta) between the direction of the secondary structure in the 3D space and the segment linking the center of that secondary structure with the RP are first calculated. (GH reference table RT)
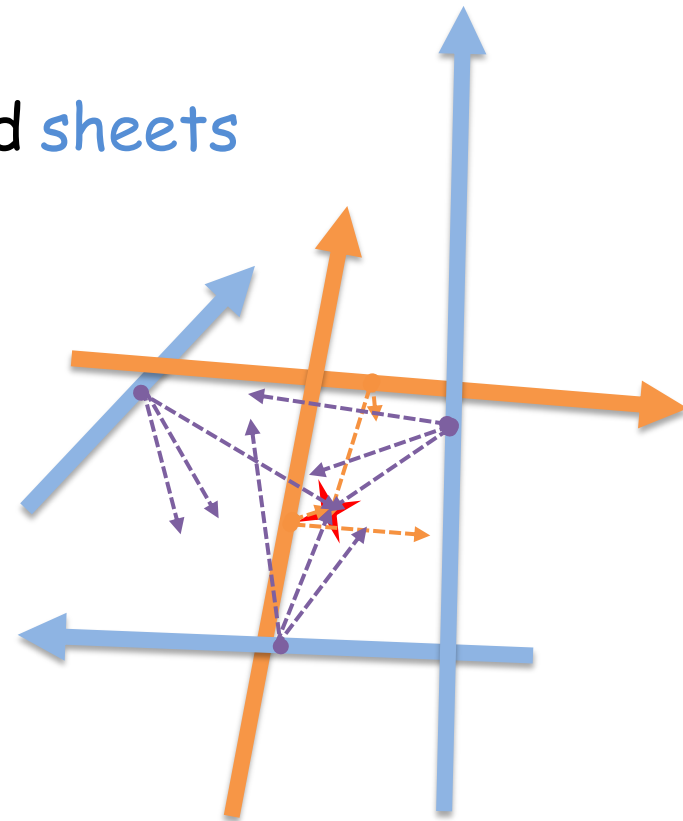
In the way of GHT
(simplified 2D representation)

helices and sheets
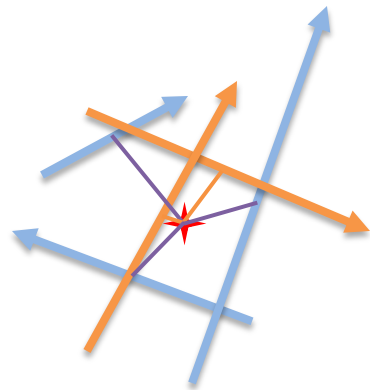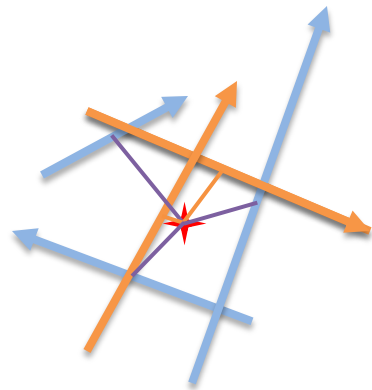

Query protein
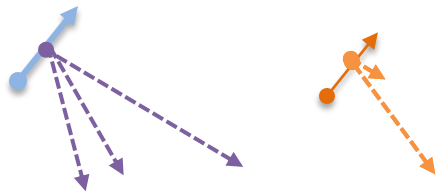(scaled 0.5)

Mapping Rule

Votes Space
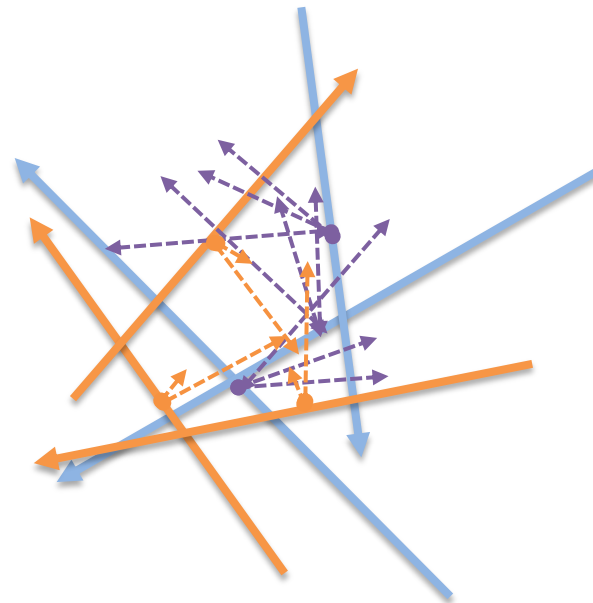
In the way of GHT

helices and sheets

Query protein
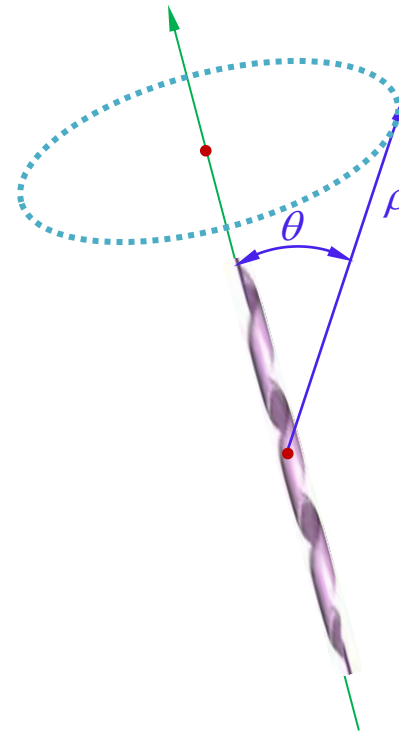
Mapping Rule

Votes Space

GHT applied to proteins

- In the 3D space of a given "object protein", every secondary structure of a "model protein" votes a circumference of points starting from every secondary structure of the object protein.

- If the proteins are similar in shape, the circumferences will all intersect in a given point.

GHT and Motif Retrieval

- Other than shape resemblance, the algorithm might be used for motif retrieval, as for instance when the "model protein" is one or many possible motifs to be looked for in a given "object protein".

Generalized Hough Transform (SSS)

Reference Point

ρ

θ

A

Type A: α-helix, l_1, TD..
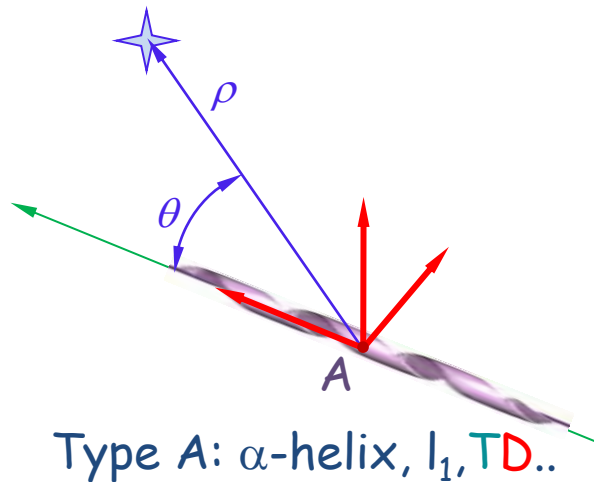
Main characteristics

- the mapping rule, for each compatible correspondent, in 3D is a circle on a plane perpendicular to the axis of the secondary structure

- Other information can be exploited to increase the S/N ratio:
  - the length of the secondary structure
  - the residues properties contained in the SS
  - any other (biochemical, morphological, etc.) peculiarities.

The implementation

- The voting space is smoothed by accumulation of nearby votes (within a given radius) for each point

- After smoothing, the highest peaks in the voting space are detected (avoiding to pick high votes that however are not the top of a peak but lie close to one such peak)

- Only the relevant votes are stored in memory: there isn't a matrix with all the possible cells.

Smoothing Algorithm

- Smoothing is performed by accumulating votes within a given radius, for every point in the vote space.
- The classic version, i.e., checking every vote for the vicinity condition, has been proven to be too time-consuming for applications, with a time complexity of $O(n^2)$, where n is the number of votes in the vote space.
- The smoothing problem can be seen as an "orthogonal search" problem, i.e., finding points within a given cube in space.
- A particular structure has been implemented for solving this problem with a $O(n \log^3(n))$ complexity: Range Trees.
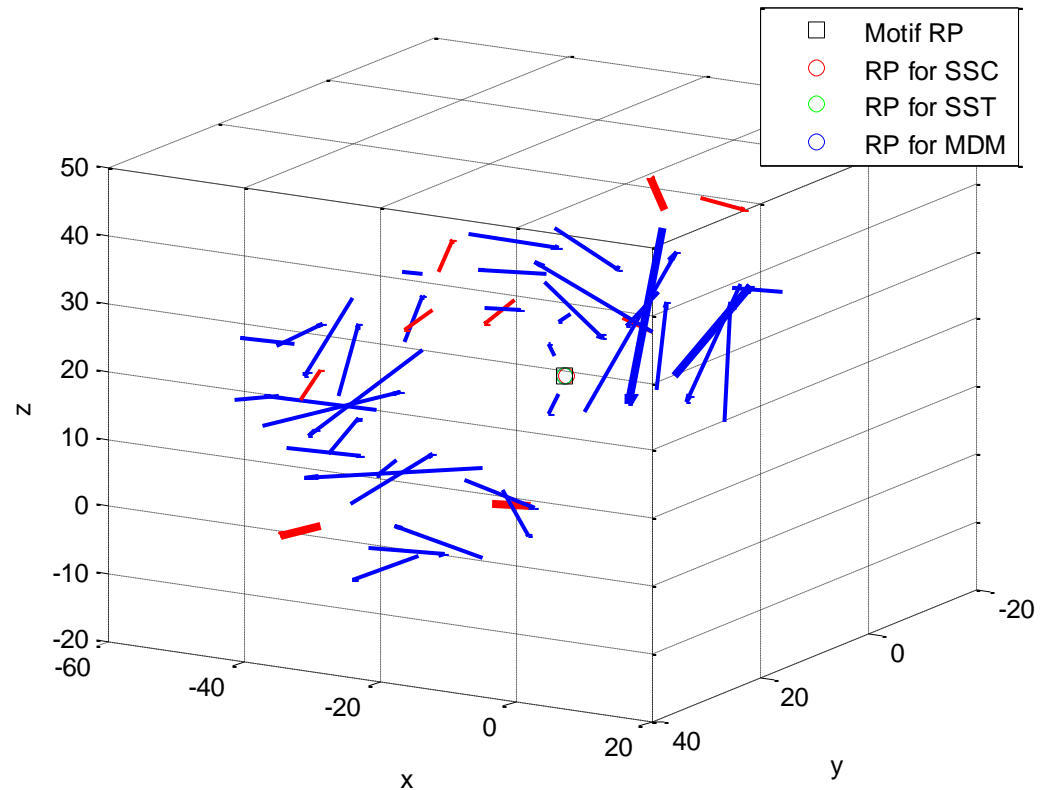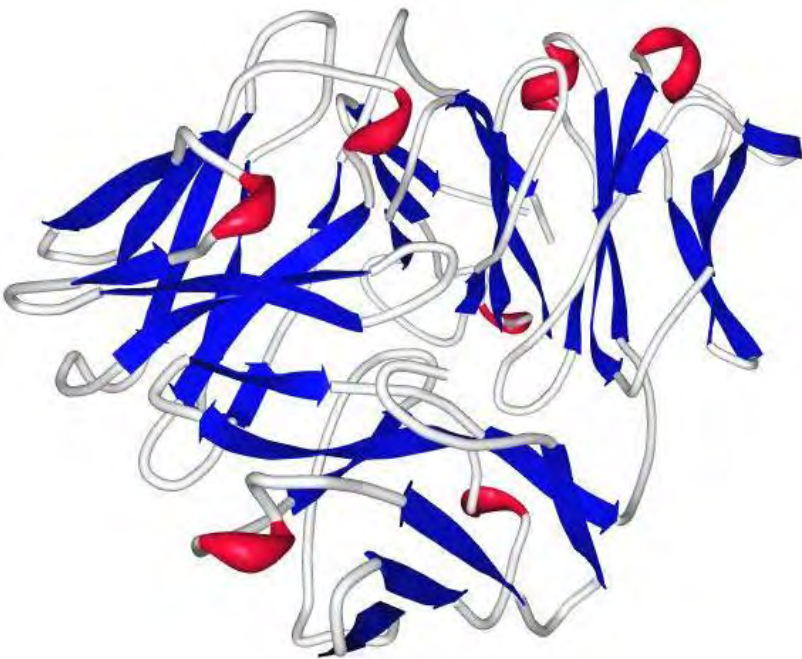
PROTEIN 1FNB
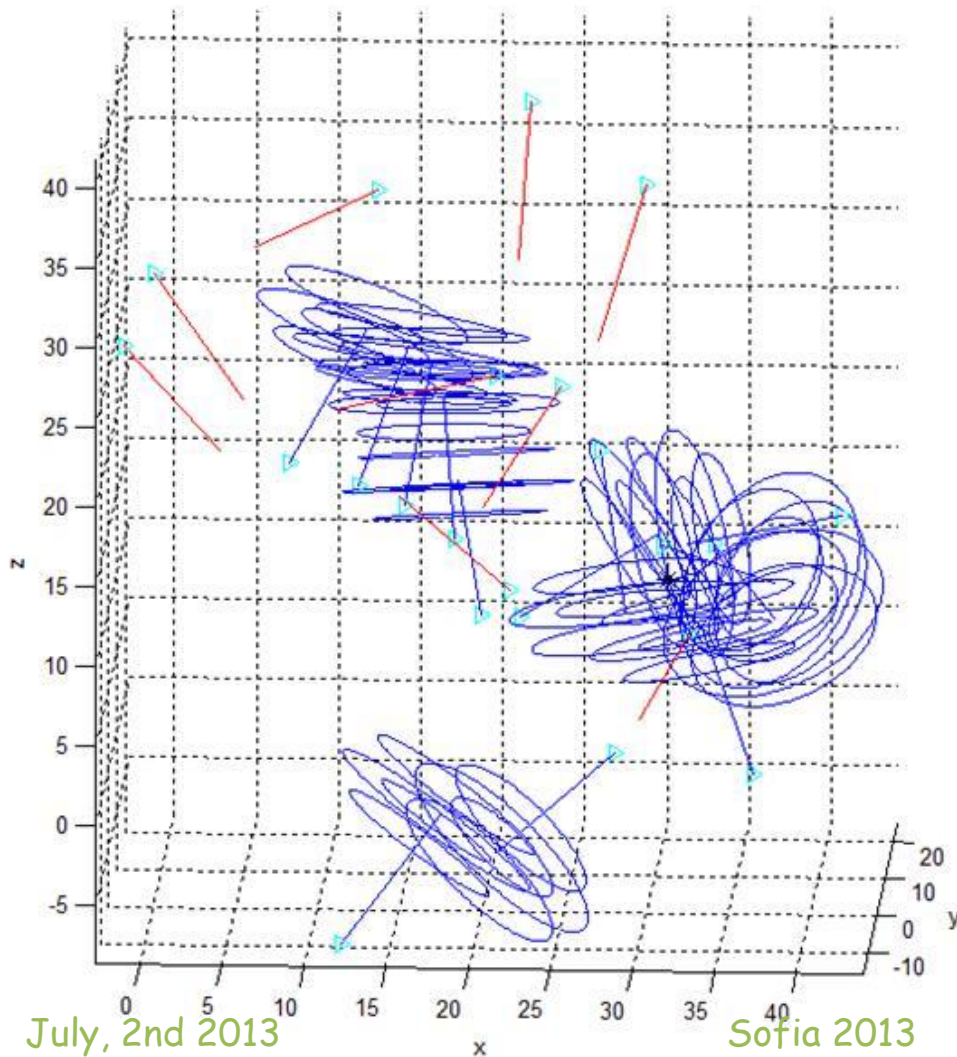
The protein contains 22 Secondary Structure.
Searched motif: Greek key (4 β-sheets).
The red circles are the *helices* and the blue circles are the *sheets*.
The cyan blue triangles indicate the *orientation* of the secondary structures.
The **black** point is the *reference point*.

Protein 7fab
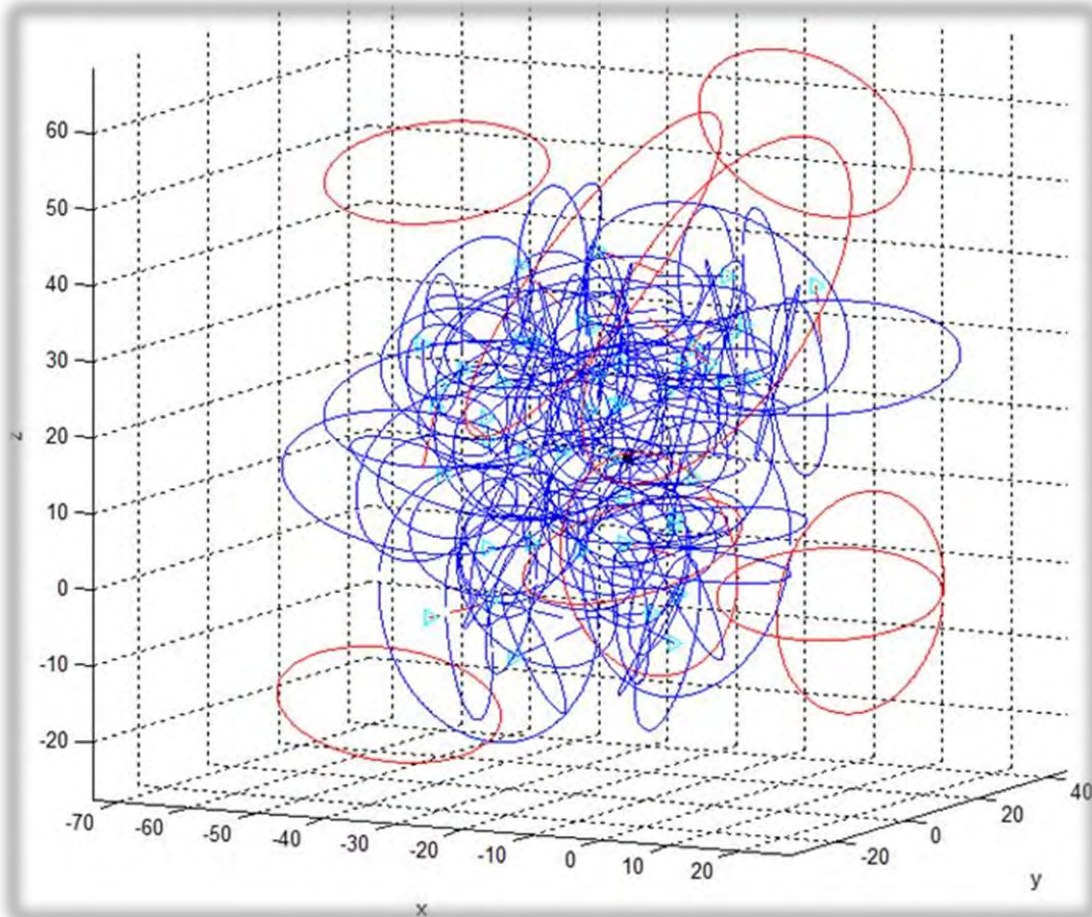
The protein contains 46 Secondary Structure.
Searched motif: 3 helix and 2 sheet.
The red circles are the *helices* and the blue circles are the *sheets*.
The cyan blue triangles indicate the *orientation* of the secondary structures.
The **black** point is the *reference point.*

SSC: Secondary Structures Co-occurrences

RP

Axis angle φ

B

Coplanar lines

Midpoint distance

Axis distance

A

Type A: π-helix, l_1, TD..
Type B: α-helix, l_2, TP..
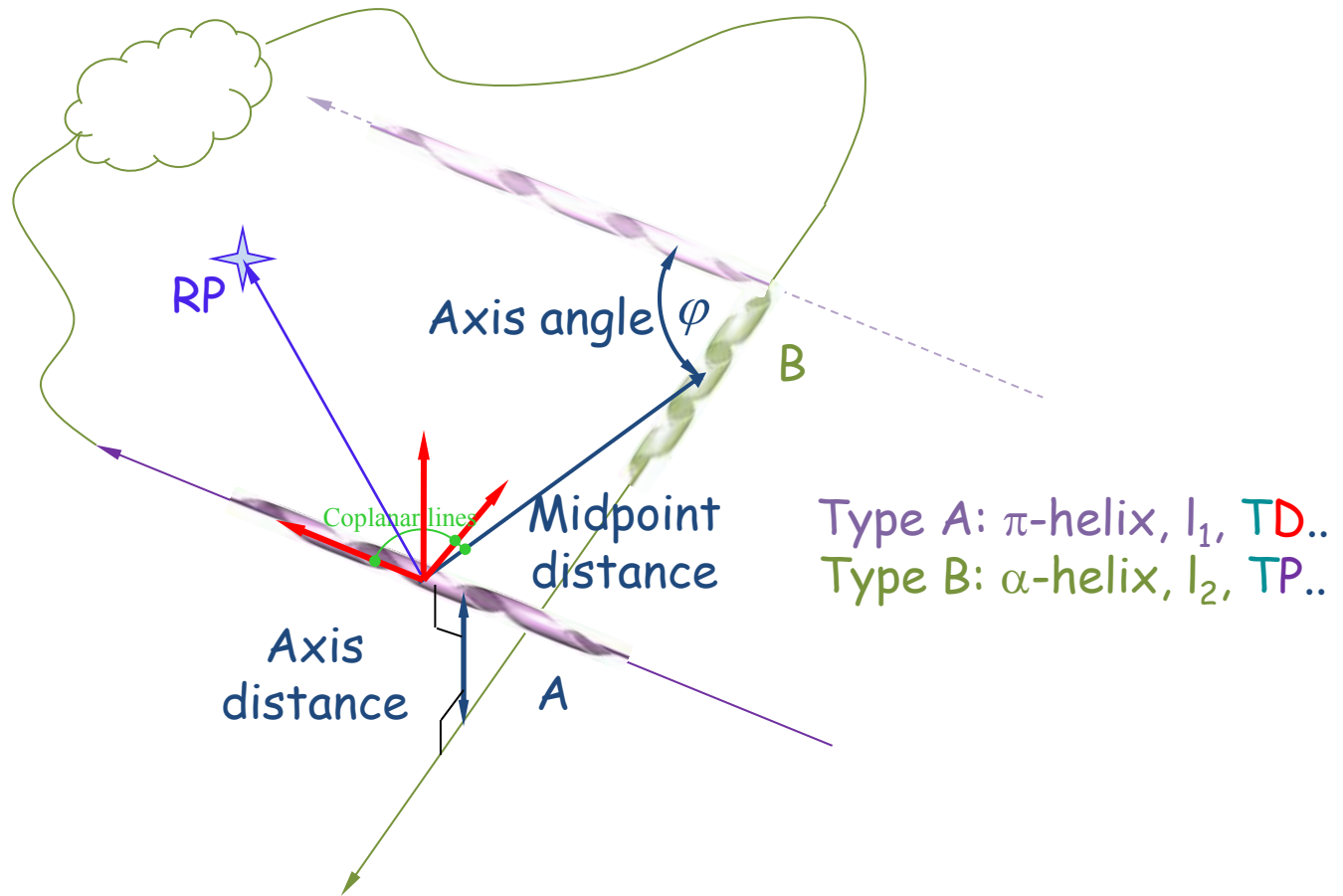
SST: Secondary Structures Triplets

Reference Point

normal to π

$\widehat{ABC} : l_{AB}, l_{BC}, l_{CA}$

C

normal to π

B

$\pi : \widehat{ABC}$

A
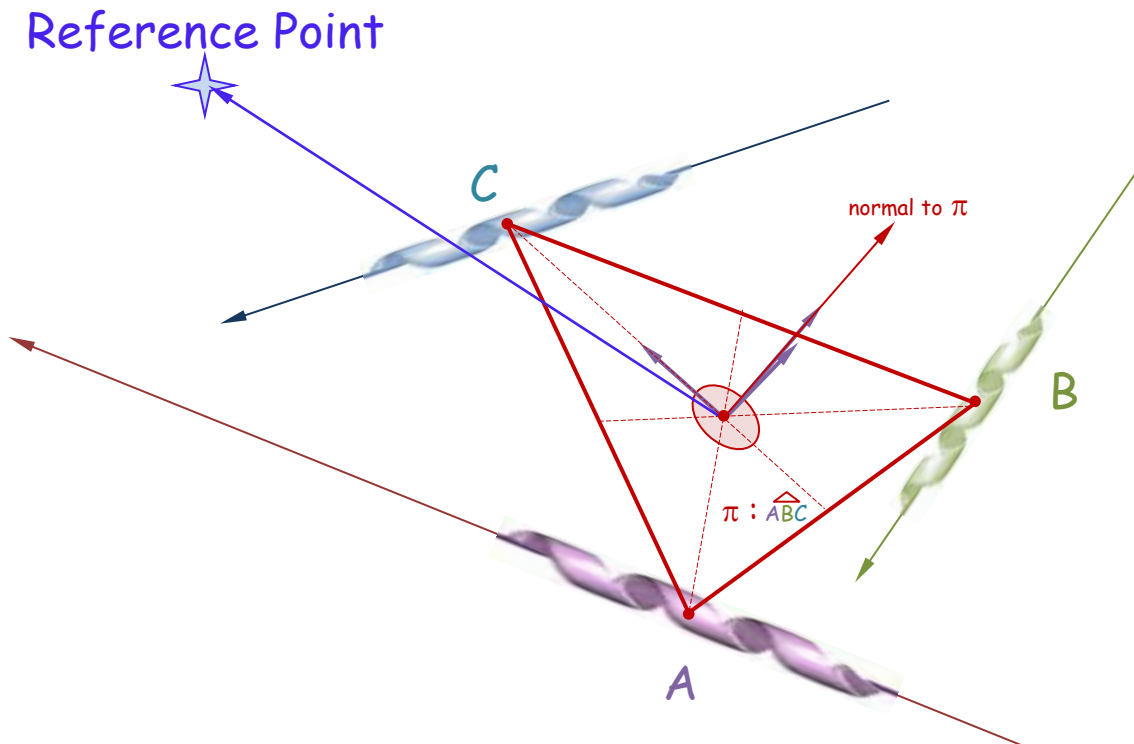
Type A: π-helix, l_1
Type B: α-helix, l_2
Type C: α-helix, l_3

4 SSs motif:
Terns co-occurrence
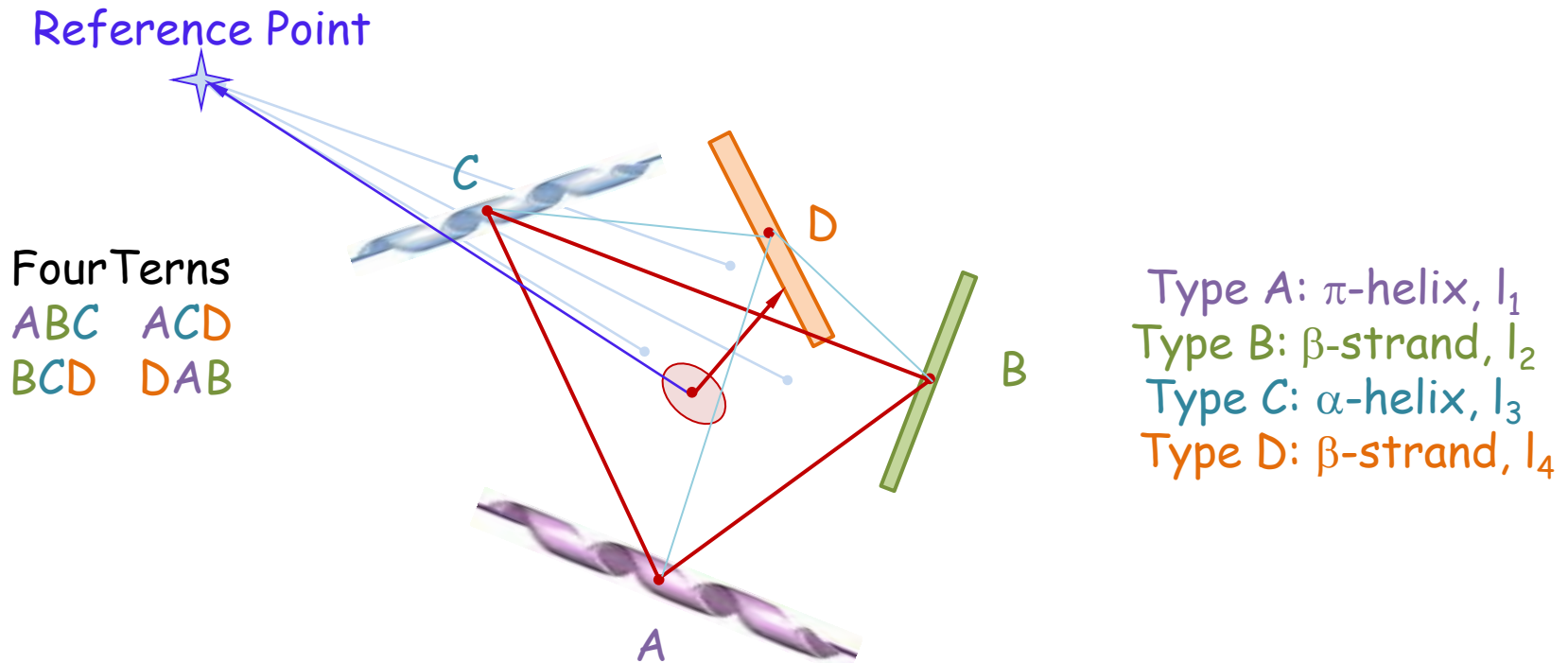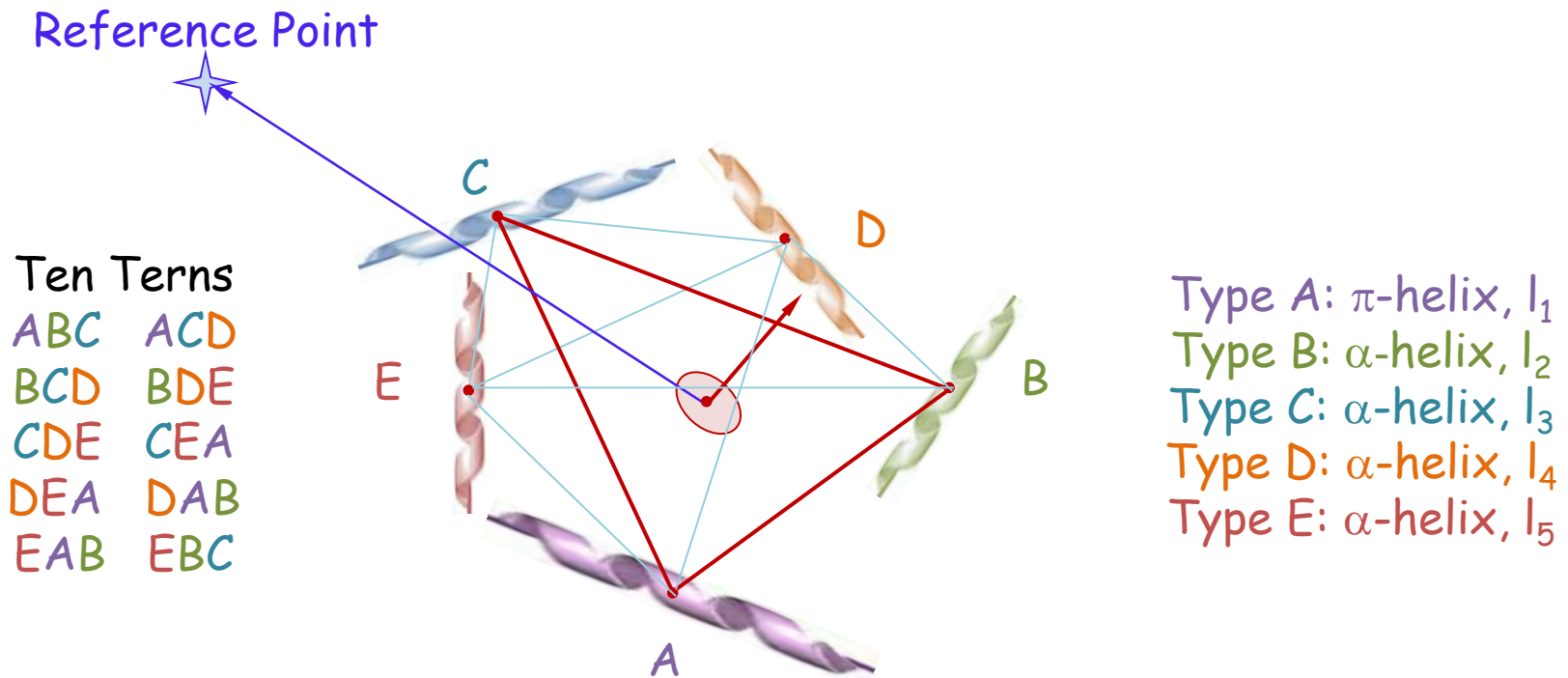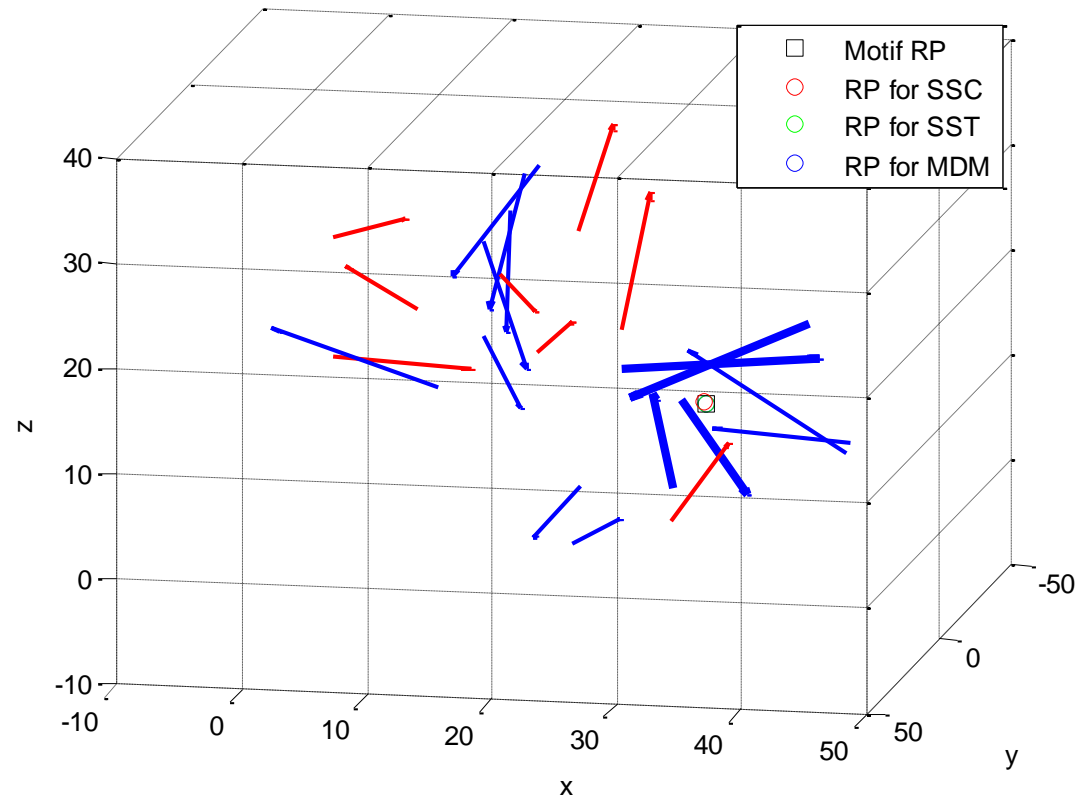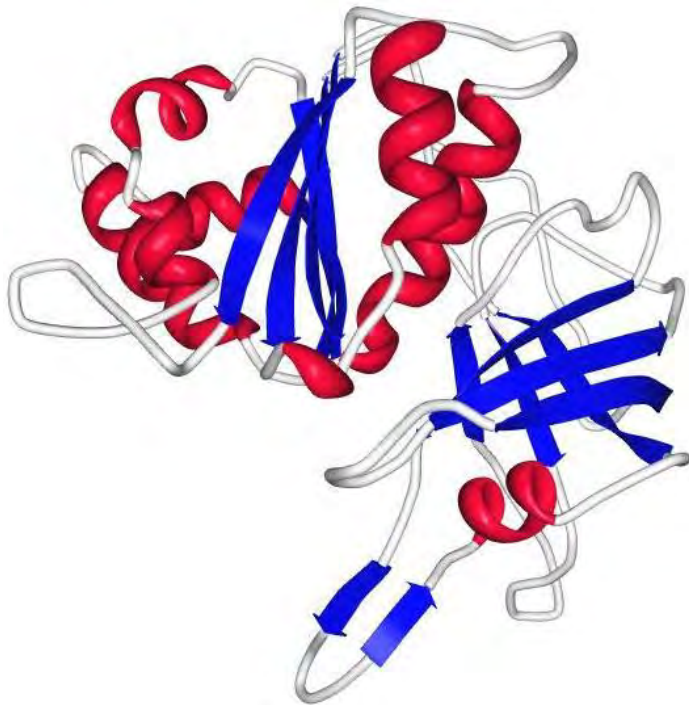
Reference Point

C

D

FourTerns
ABC   ACD
BCD   DAB

Type A: π-helix, l_1
Type B: β-strand, l_2
Type C: α-helix, l_3
Type D: β-strand, l_4

B

A

5 SSs motif:
Terns co-occurrence

Reference Point

Ten Terns

ABC   ACD
BCD   BDE
CDE   CEA
DEA   DAB
EAB   EBC

C

D

E

B

A

Type A: π-helix, l_1
Type B: α-helix, l_2
Type C: α-helix, l_3
Type D: α-helix, l_4
Type E: α-helix, l_5

PROTEIN 1FNB

The protein contains 22 Secondary Structure. Searched motif: Greek key (4 β-sheets). The red circles are the *helices* and the blue circles are the *sheets,* in **bold** the motif SSs.

PROTEIN 7FAB

The protein contains 46 Secondary Structure. Searched motif: 3 helices and 2 sheets.
The red circles are the *helices* and the blue circles are the *sheets,* in **bold** the motif SSs.

Searching performances

Searching a Greek Key motif (4 SSs, all β-sheets) in 1FNB

Methods	Candidate RP	Error Rate %	SearchTime
SSS	[31.41 1.16 11.94]	0.32	35.2sec
SSC	[31.33 1.08 11.79]	0.33	3.86sec
SST	[31.38 1.08 11.69]	0.00	5.76 sec
MDM	[31.40 1.12 11.66]	0.16	7.59sec

Searching a motif with 5 SSs (3 helices and 2 sheets) in 7FAB

Methods	Candidate RP	Error Rate %	Search Time
SSS	[-17.56 9.46 15.17]	0.28	108sec
SSC	[-17.40 9.14 15.48]	0.34	42.51sec
SST	[-17.48 9.17 15.48]	0.00	48.89sec
MDM	[-17.45 9.16 15.50]	0.15	112.17sec

PV Benchmark (20 proteins)

PDB ID	Description	Number of SSs
2Z98	FMN-dependent NADH-azoreductase	14
2Z9C	FMN-dependent NADH-azoreductase	15
2Z9B	FMN-dependent NADH-azoreductase	16
4GCR	GAMMA-B CRYSTALLIN	18
3E9O	Pre-mRNA-splicing factor 8	21
1FNB	FERREDOXIN-NADP+ REDUCTASE	22
3E9L	Pre-mRNA-processing-splicing factor 8	23
2PZN	Aldose reductase	24
3C3U	Aldo-keto reductase family 1 member C1	26
2Z7G	Adenosine deaminase	28
2Z7K	Queuine tRNA-ribosyltransferase	33
2PRL	Dihydroorotate dehydrogenase, mitochondrial	34
2QX8	Ribosyldihydronicotinamide dehydrogenase [quinone]	35
2QMY	Ribosyldihydronicotinamide dehydrogenase	36
3C94	Exodeoxyribonuclease I	37
2QX9	Ribosyldihydronicotinamide dehydrogenase [quinone]	38
3C95	Exodeoxyribonuclease I	39
3DC7	Putative uncharacterized protein lp_3323	43
3DHP	Alpha-amylase 1	44
7FAB	IGG1-LAMBDA NEW FAB (LIGHT CHAIN)	46

PV Benchmark: basic features
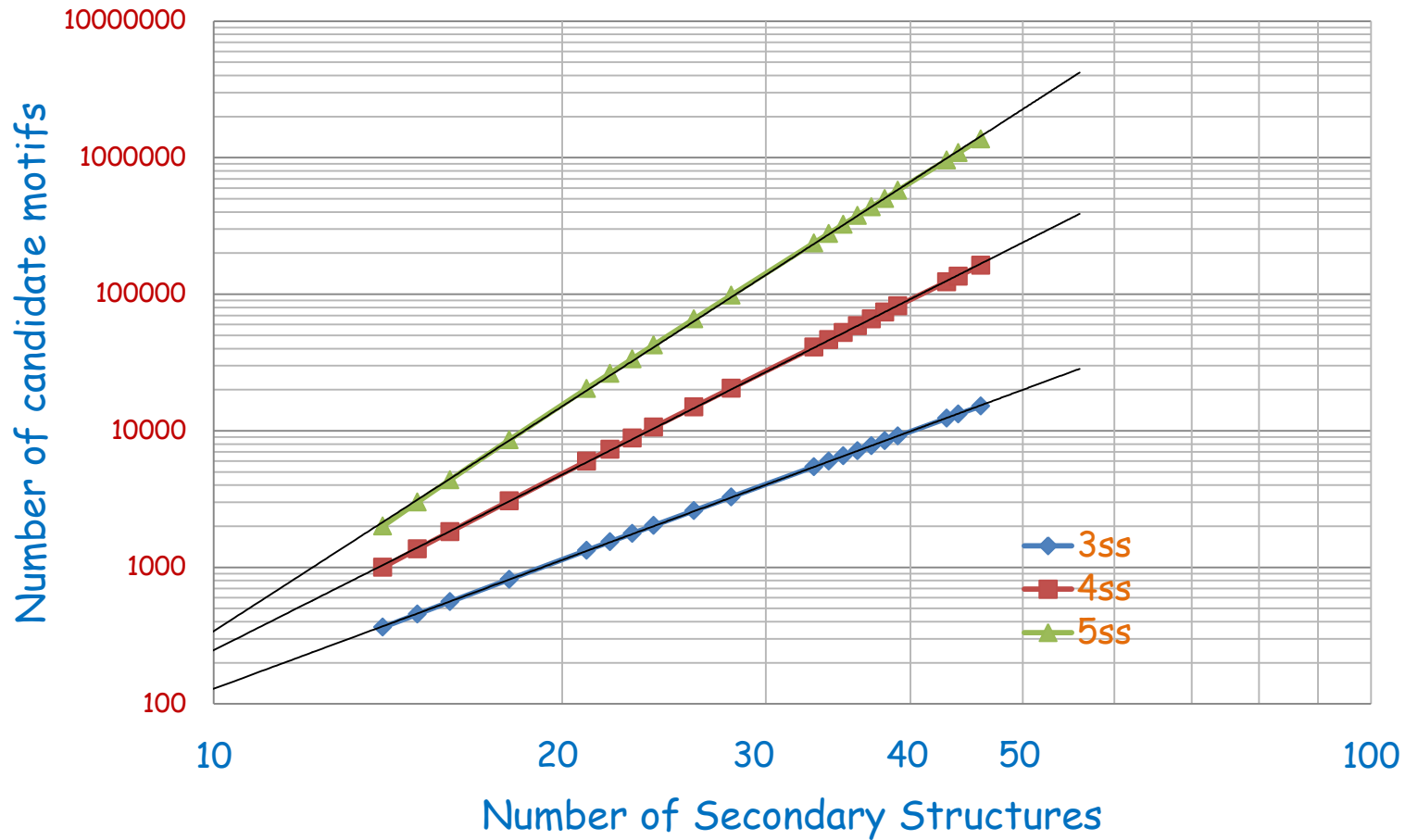
PV Benchmark: performances
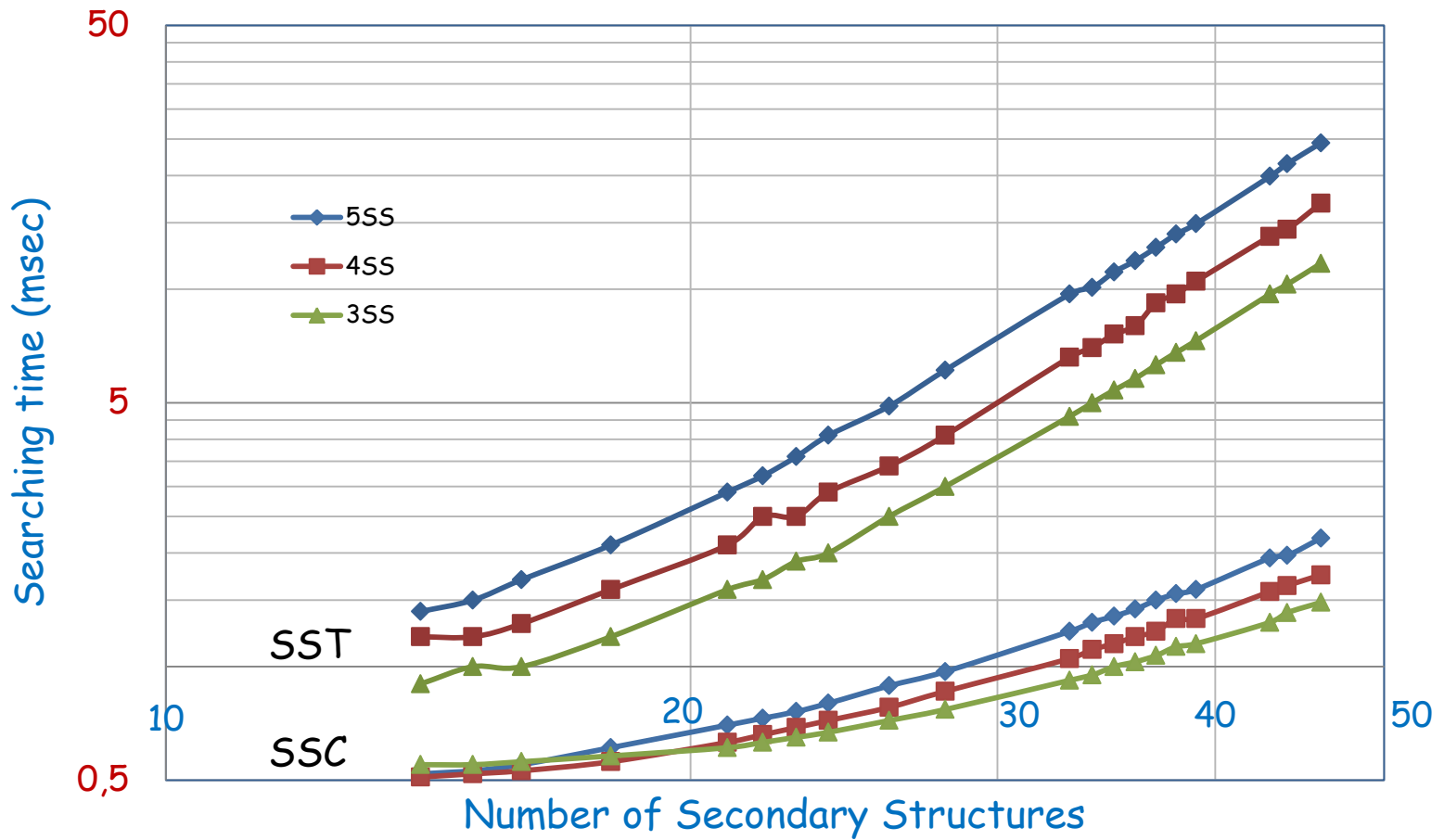
Average performances

SSC

Number of proteins	Number of SSs per motif	Total number of motifs	Total Searching Time (sec)	Average Searching Time per motif (msec)
20	3	105971	119.882	1.1 [0.55-1.48]
20	4	918470	1275.585	1.4 [0.51-1.75]
20	5	6455009	11261.911	1.7 [0.52-2.19]

SST

Number of proteins	Number of SSs per motif	Total number of motifs	Total Searching Time (sec)	Average Searching Time per motif (msec)
20	3	105971	768.508	7.3 [0.9-11.7]
20	4	918470	10303.806	11.2 [1.2-16.9]
20	5	6455009	111809.428	17.3 [1.4-24.4]

EGI - protein 1aa01
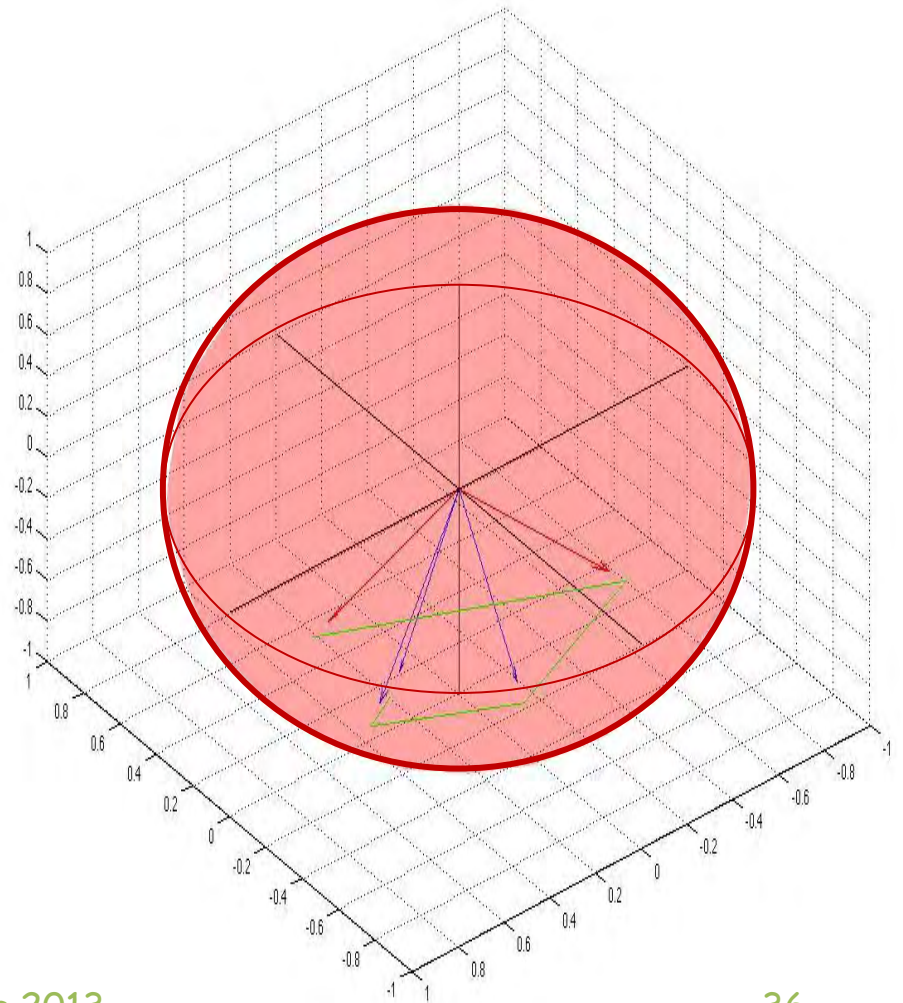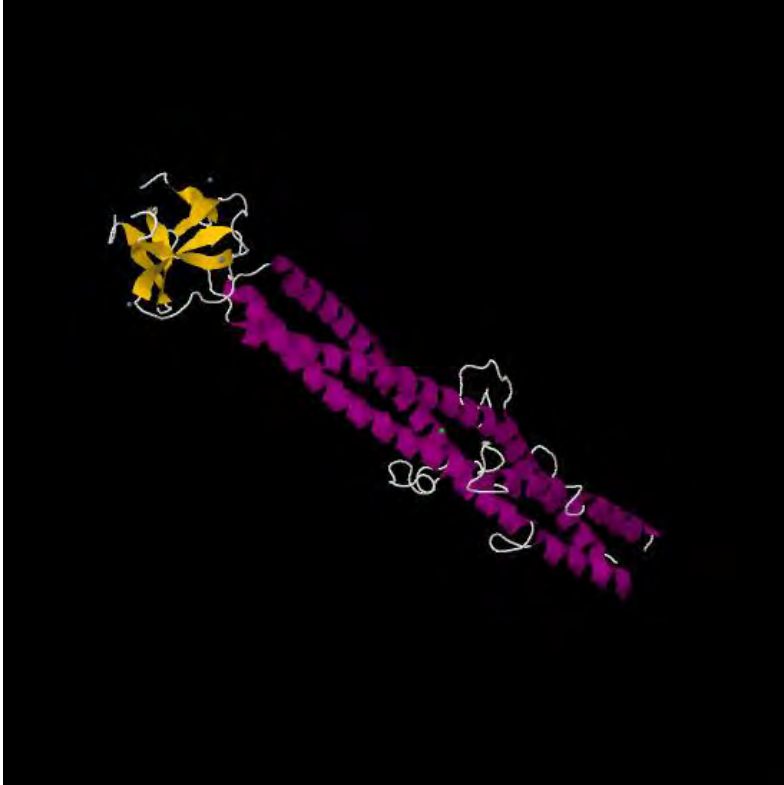
Future outlook

- These algorithms are extremely interesting, not to the least for its capability to be executed on parallel computing systems (thanks to the inherent parallel nature thereof).

- The accuracy of the comparison has the potential to be increased to competitive standards with respect to other approaches.

Example of match with cardinality 10 between
proteins 2qx8 (source) and 2qx9 (search)
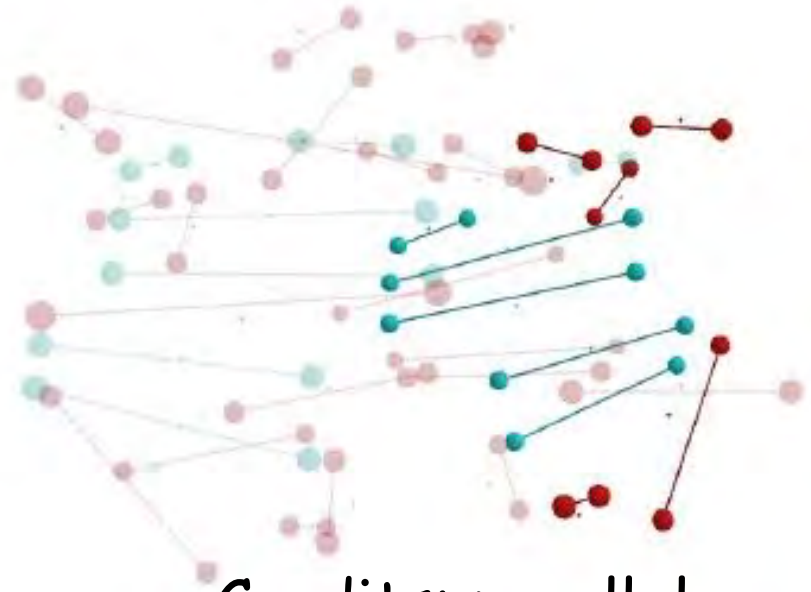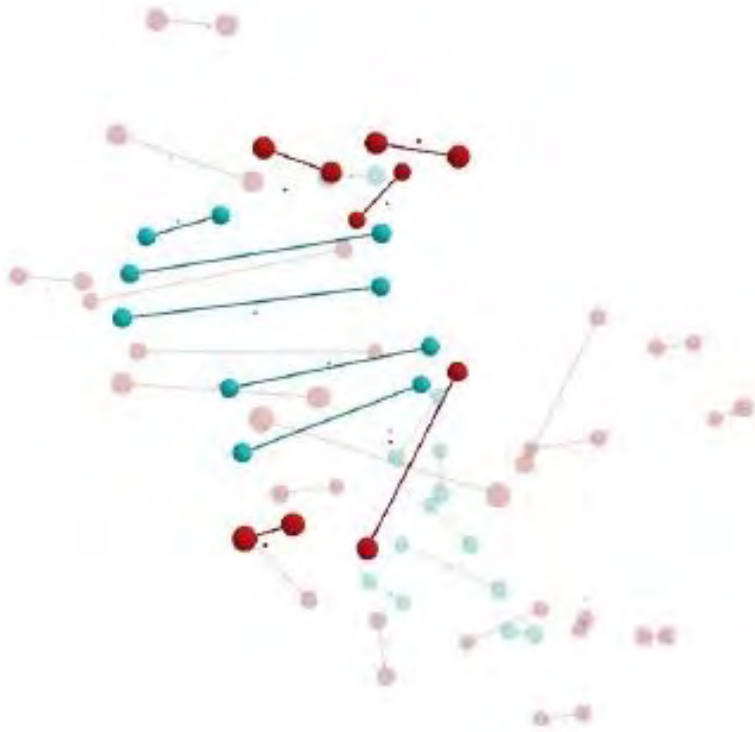
Credits: parallel
implementation
by Mirto Musci

Study interaction between
structures and other molecules
(Protein Docking)
through mathematical morphology

Surface Modeling

Van der Waals surface

Convex hull

Solvent-excluded surface

Solvent accessibile surface
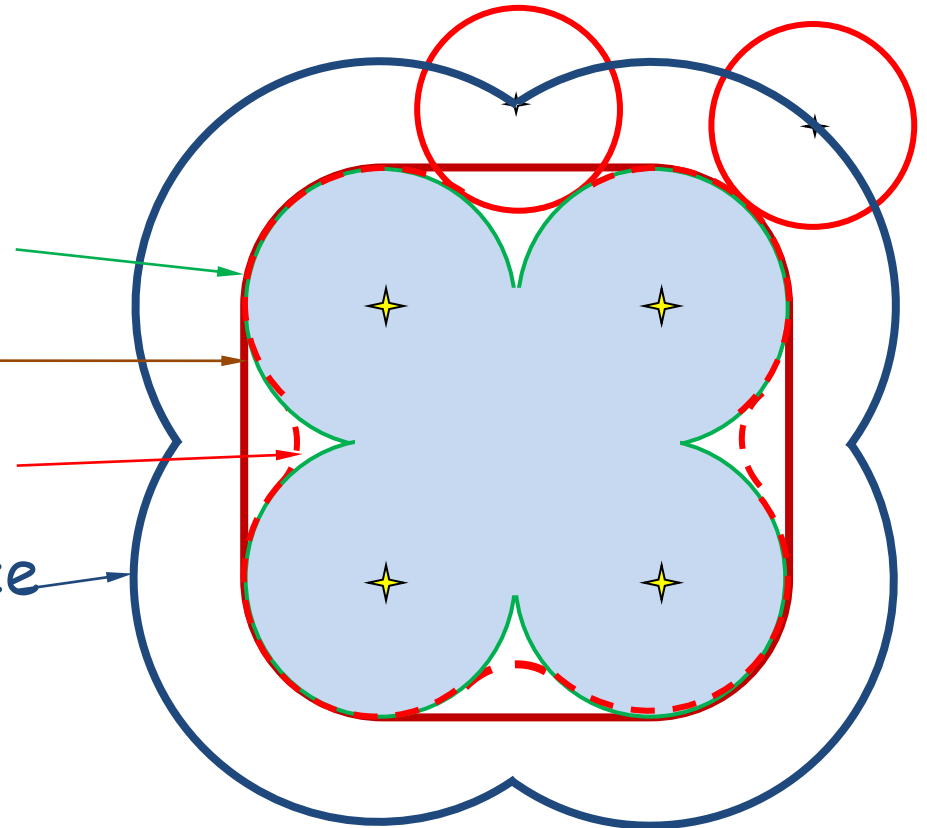
Mathermatical morphology: Closing Operation

Dilation

Erosion

The structural element is a sphere
representing the solvent
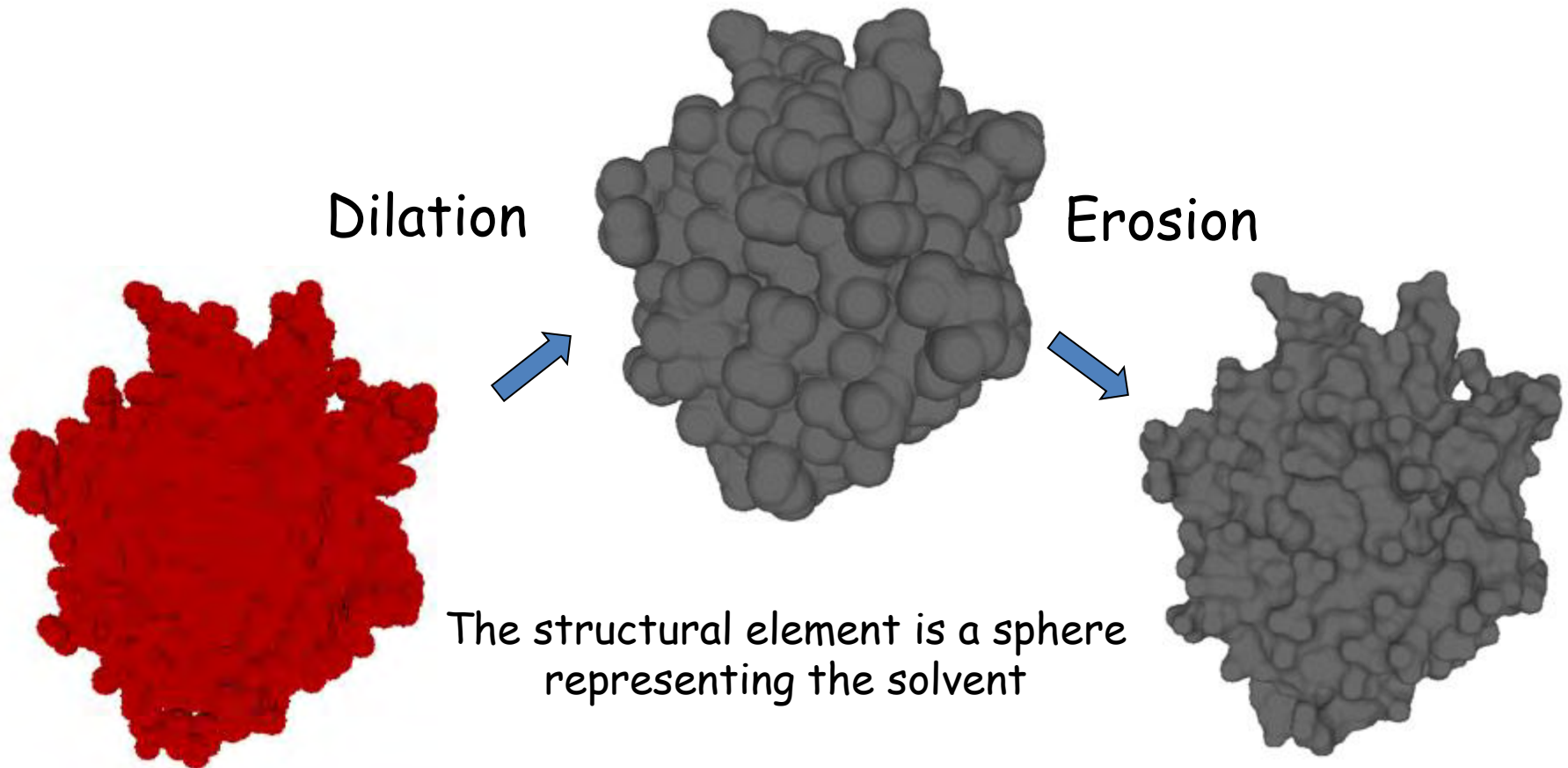
Protein-ligand docking

- A large molecule (receptor) and a small molecule (ligand) docking in a cavity.
- Key in lock
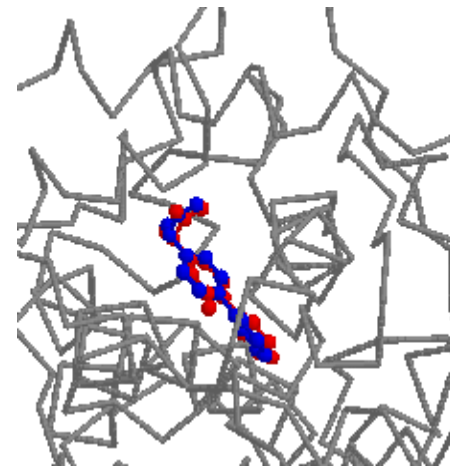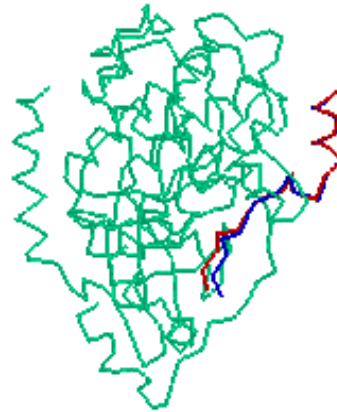
Docking
Protein-Protein Interfaces

The docking site is more planar rather than a cavity.
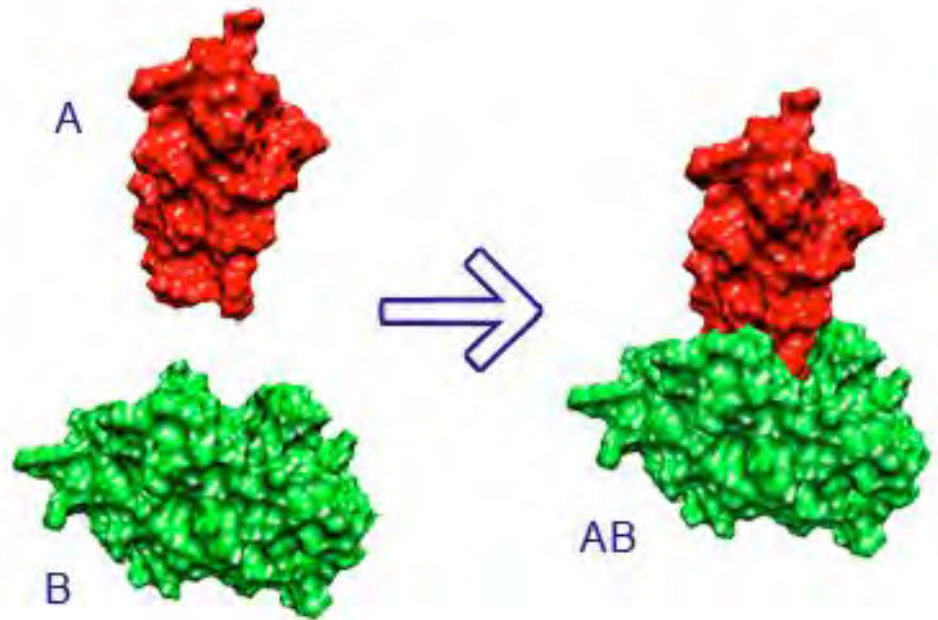
Accuracy vs coverage

- Accuracy: how many of the solutions found were correct?

  A= (F ∩ T) /F

- Coverage: How many of the correct solutions were found?

  C= (F ∩ T) /T

T: correct solutions

F:solutions found
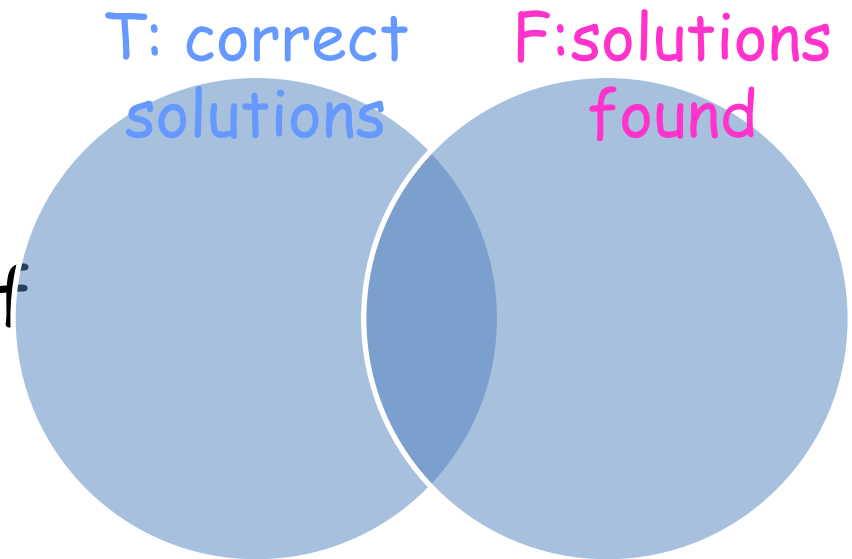
Convex Hull
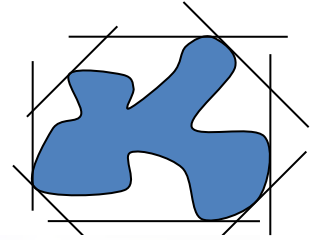
Note that straight lines (facet in 3D) are represented in a point in the EGI, and their 'spike' can be easily eliminated, leaving only the true contour (surface) of the object (protein).
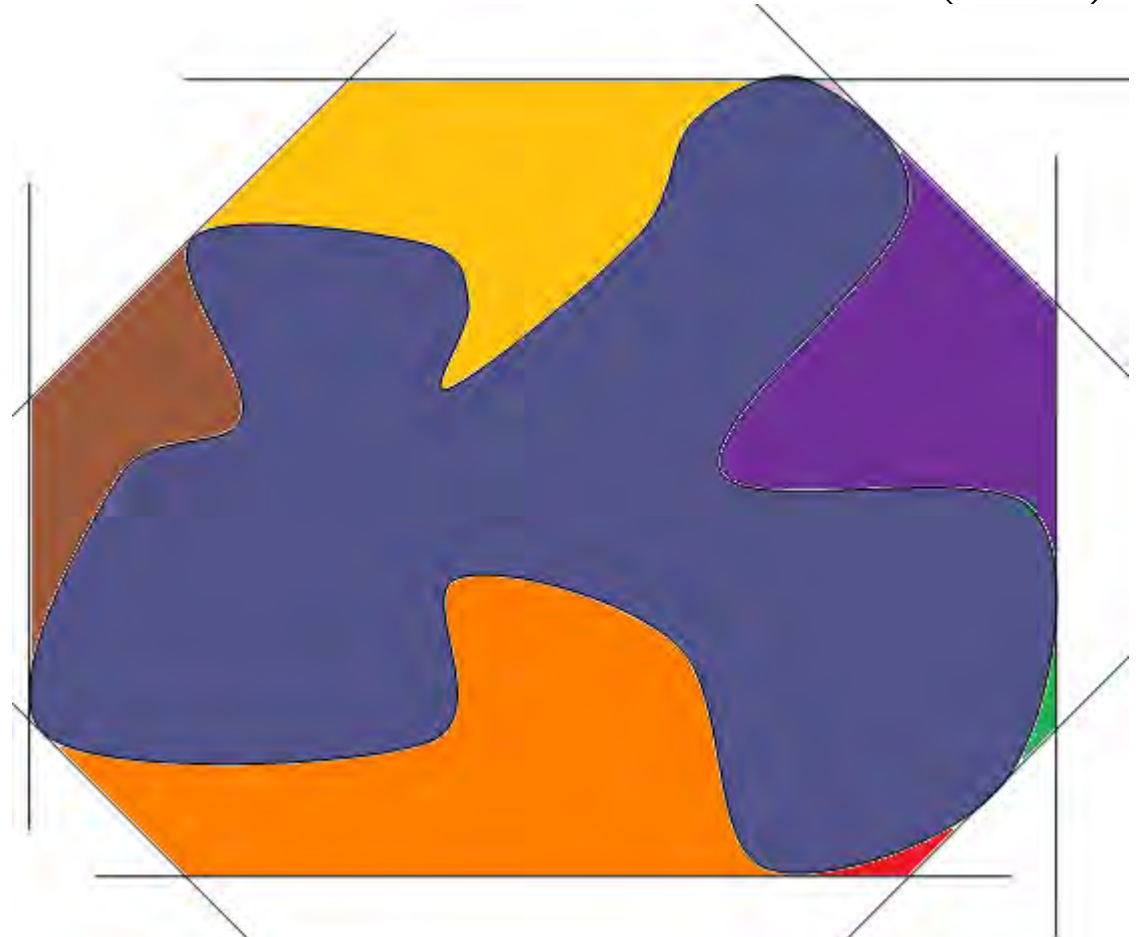
A CH of a molecule (1MK5)

- The CH of a molecule is the smallest convex polyhedron that contains the molecule

- A practical *O(n log n)* algorithm for general dimensions CH computing, is Quickhull (Barber, 1996)
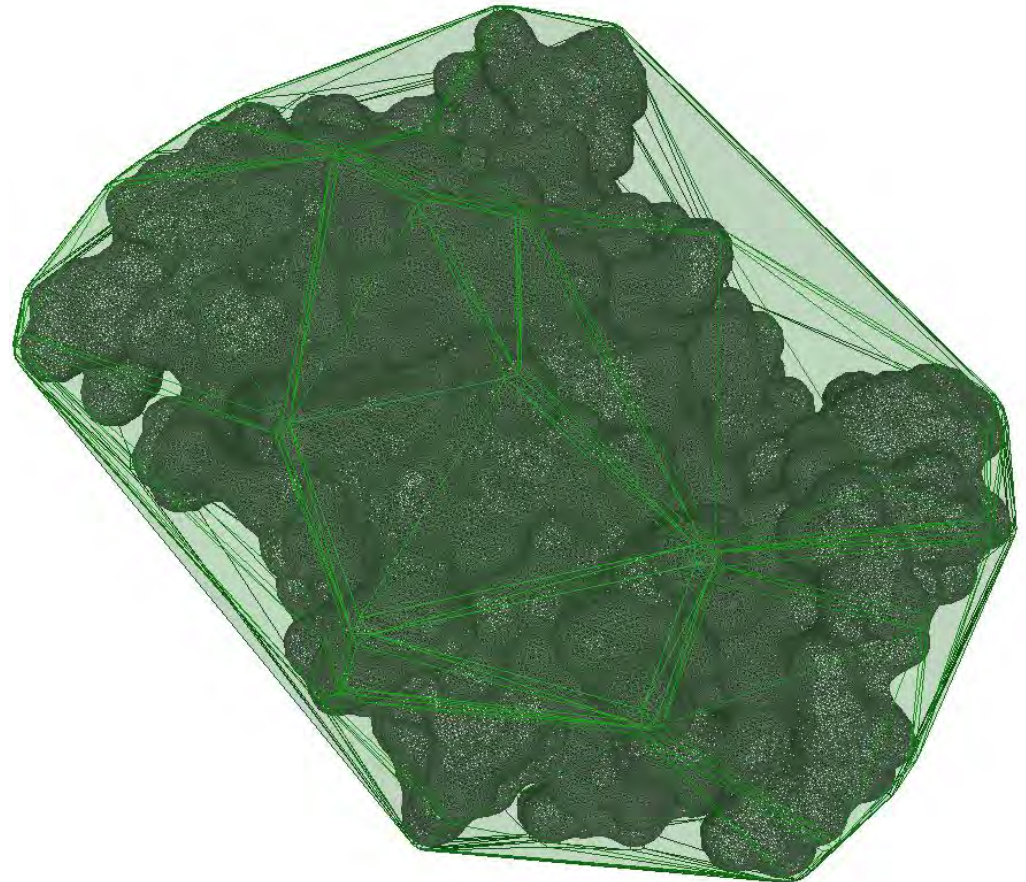
Propagation in concavity volume from CH
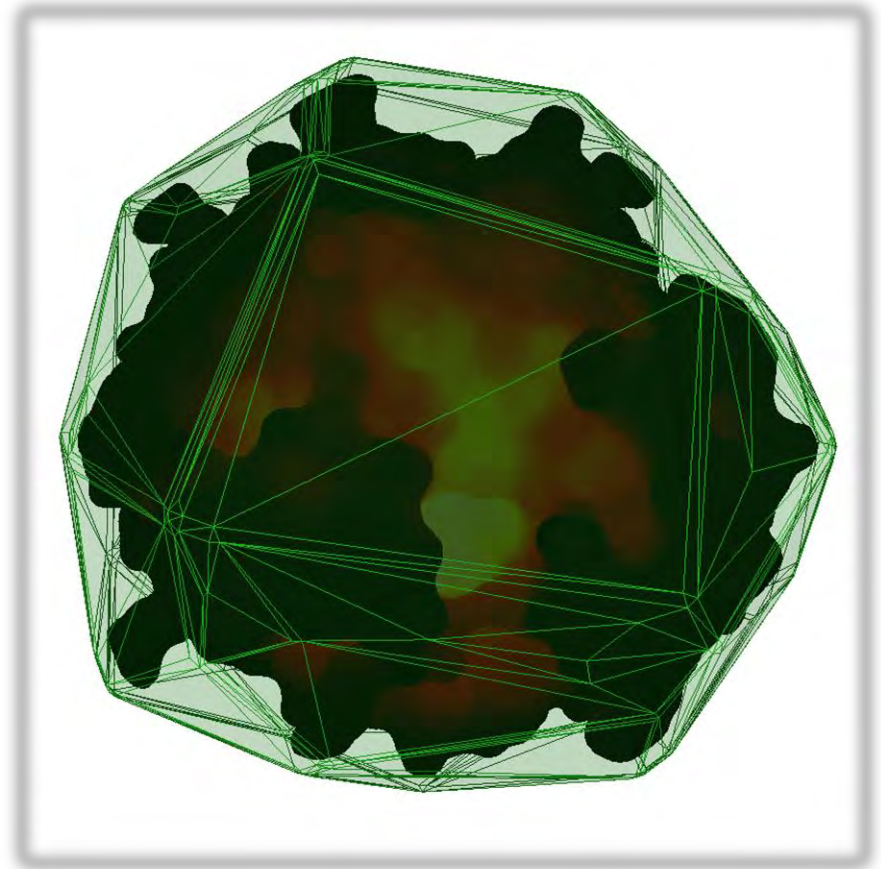
Propagation in the concavitiy volume

- The *concavity volume* is the region R between the CH and the SES
- B_{CH} is the set of the border voxels of CH
- A is the increasing set of voxels contained in R; E is to the recruited near neighbors
- $d_n + w_n$ is the minimum distance in the near neighbors by the displacement w

$$D_i = \left\{ \begin{array}{l} 1 \; iff \; i \in B_{CH} \\ 0 \; otherwise \end{array} \right\}$$

$A = B_{CH};$

$N = (A \oplus K) \cap R;$

$E = N - A;$

while $E \neq \varnothing$ do

$\quad \forall e \in E: d_e = \min_{n \in nn_e} (d_n + w_n);$

$\quad A = N;$

$\quad N = (A \oplus K) \cap R;$

$\quad E = N - A;$

done

Propagation in the CV: 2D sketch
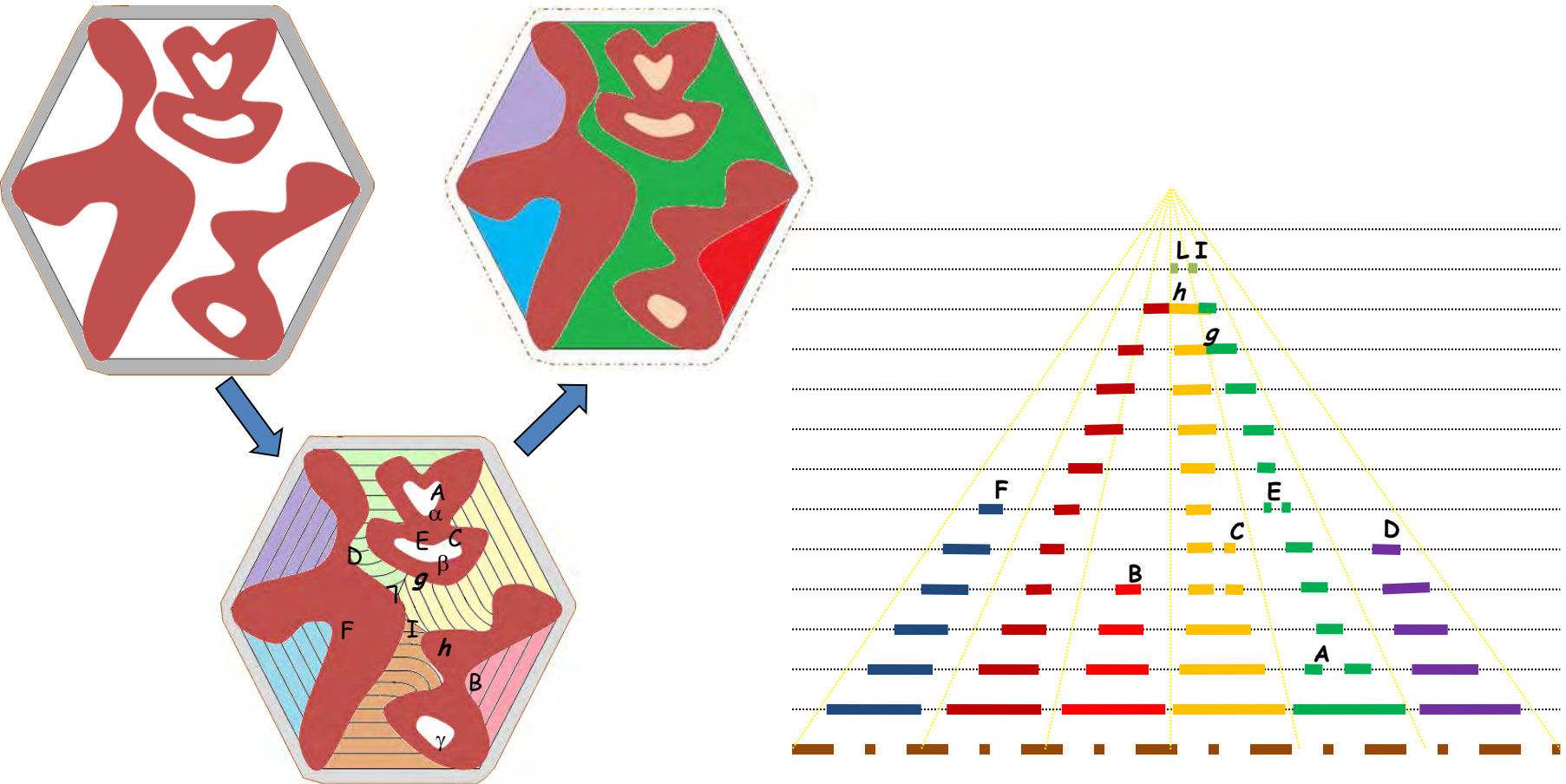
Back-propagation for pockets and tunnels search (1MK5)
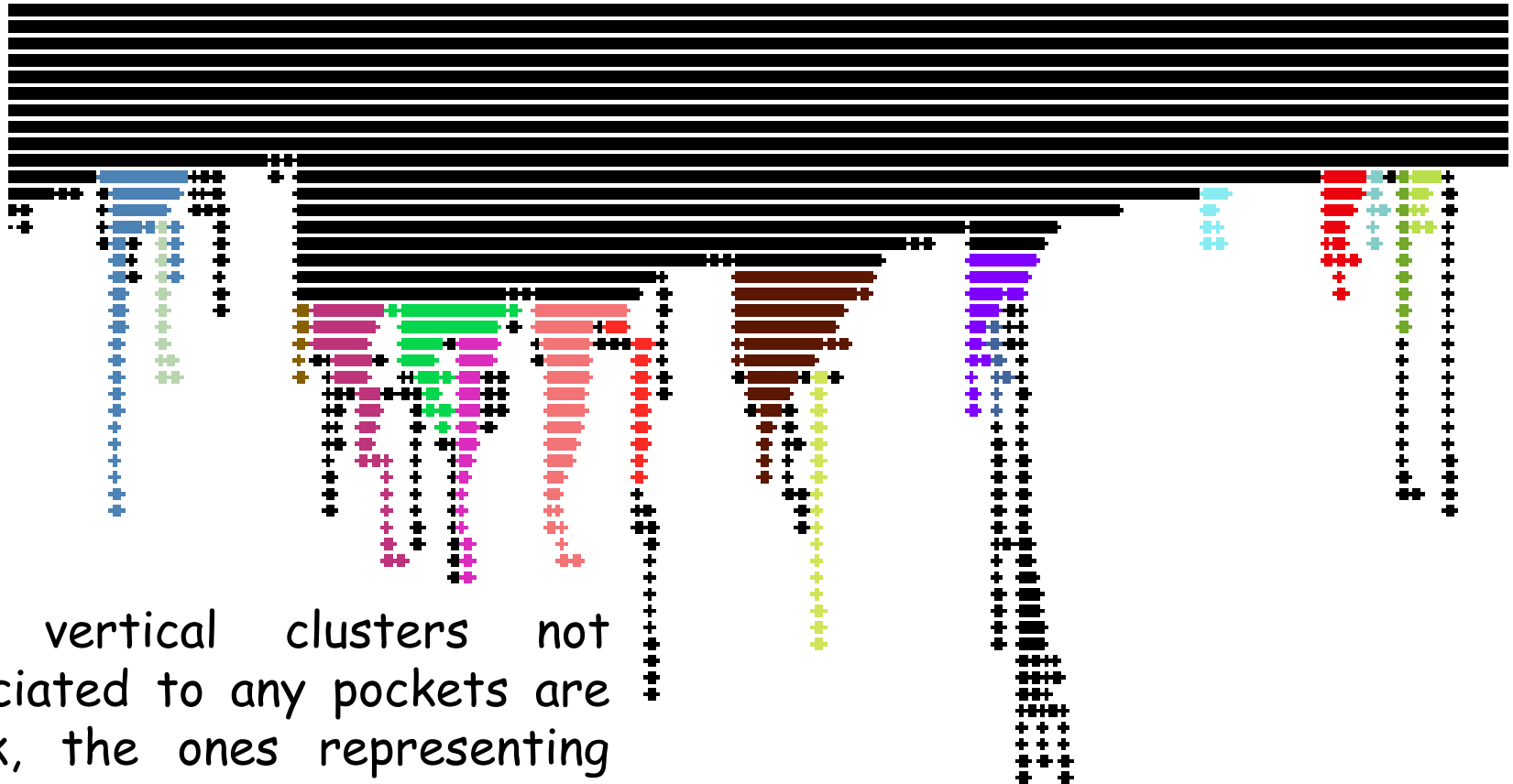
The vertical clusters not associated to any pockets are black, the ones representing pockets are colored

Pockets detected on the 1MK5
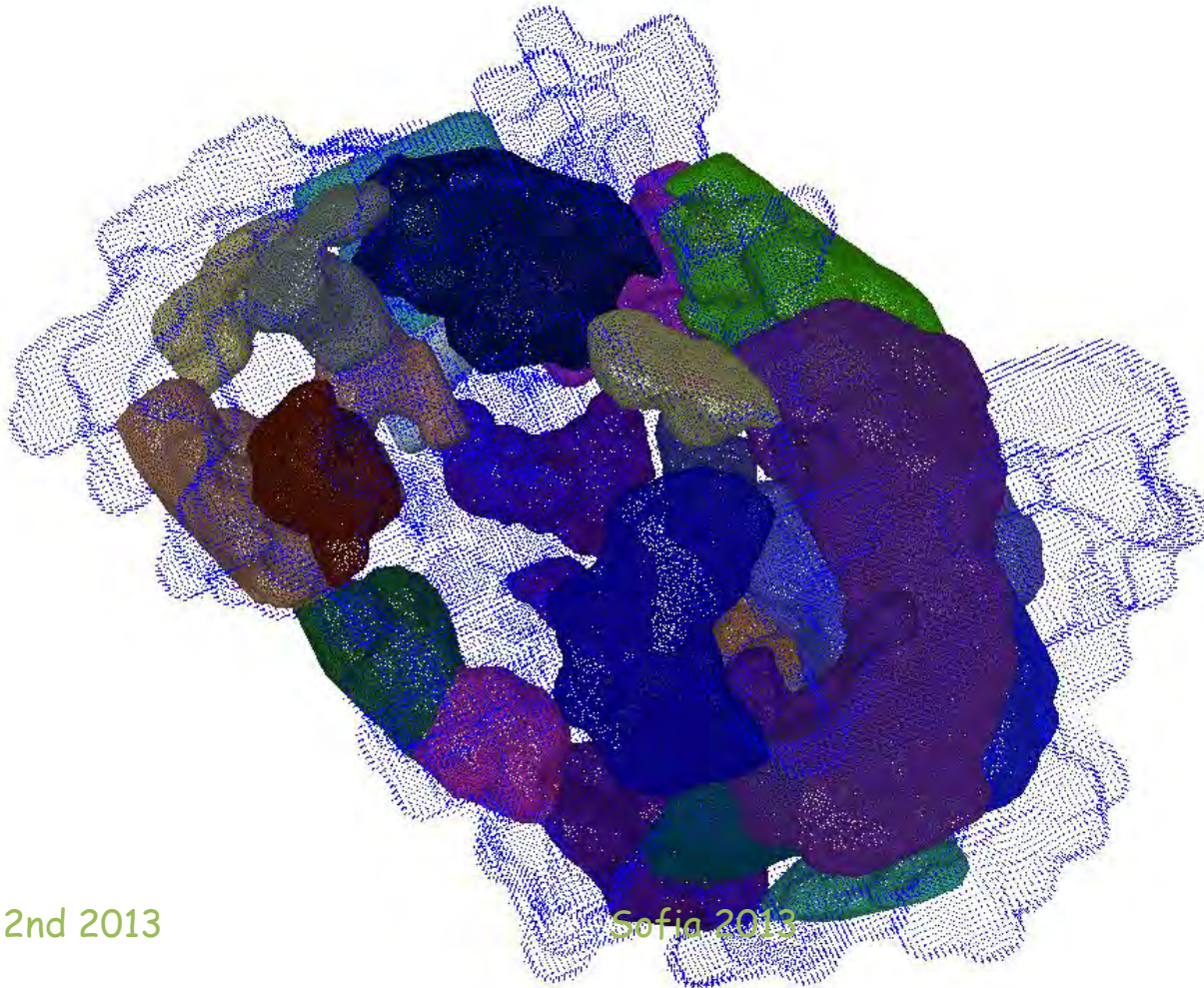
Protein Inspector – finding pockets
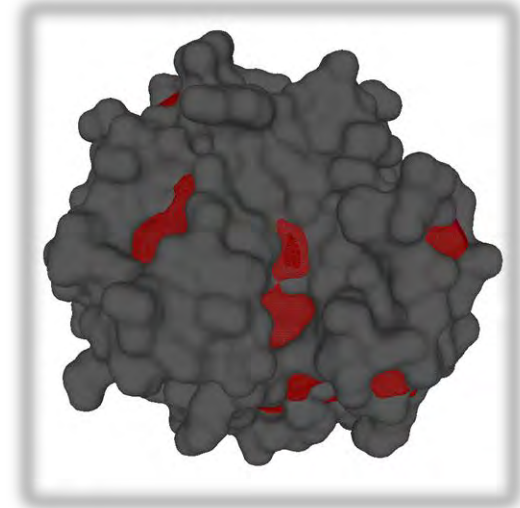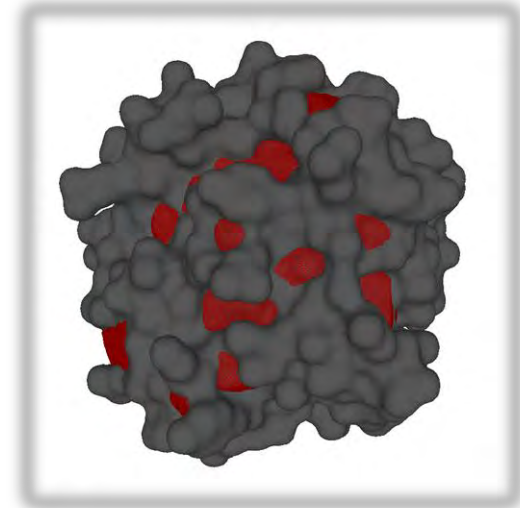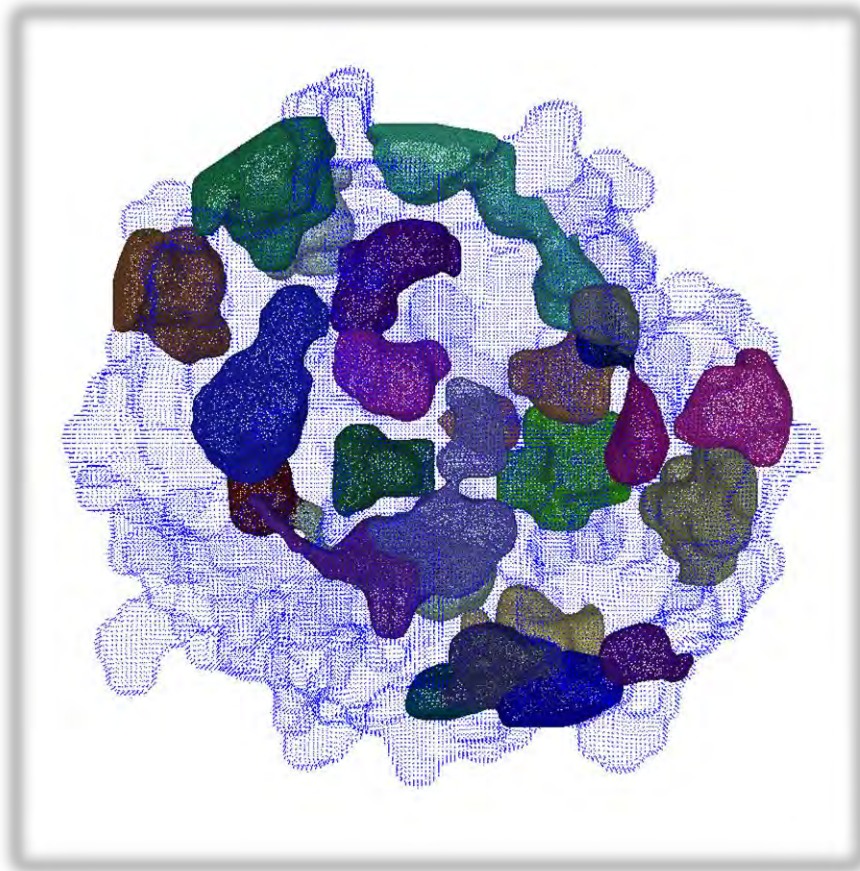
Protein Inspector: pockets refinement

All pockets at once with TD constraint
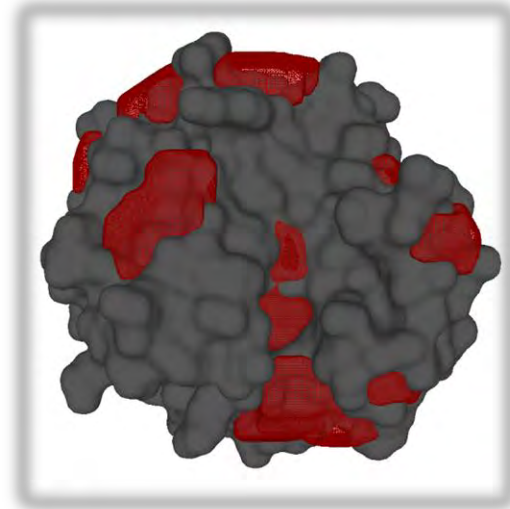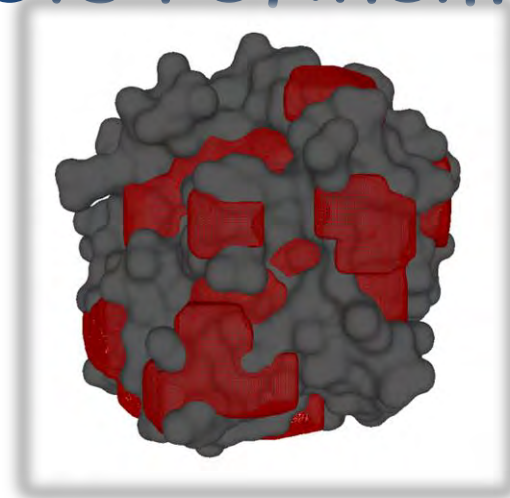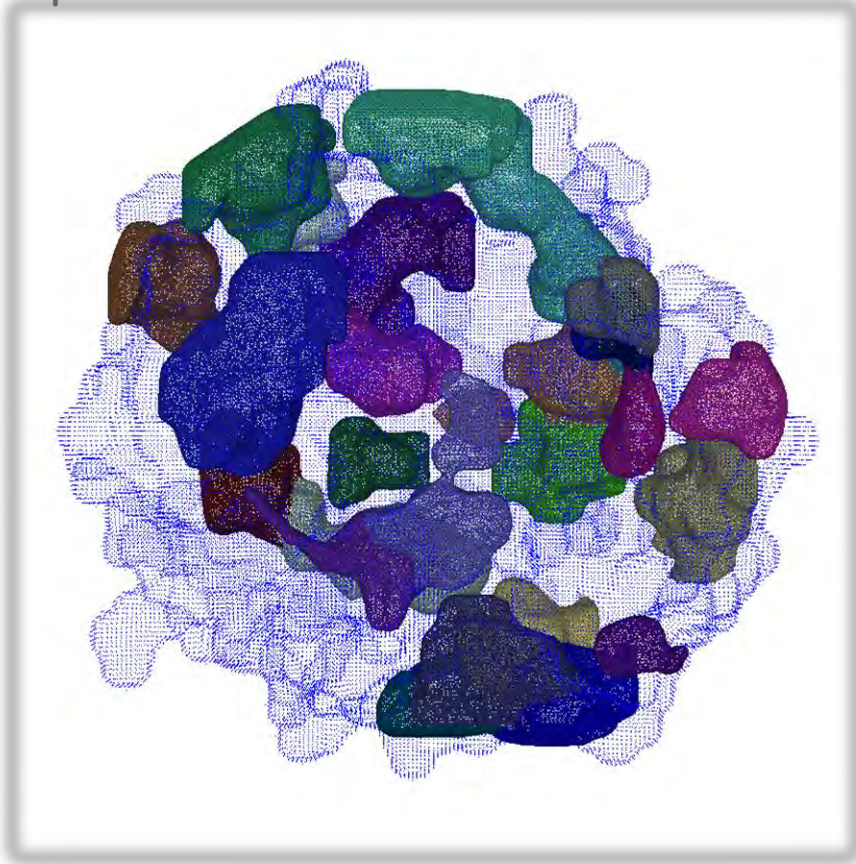
Protein Inspector – pockets refinement

Single pocket with TD and ligand box constraints
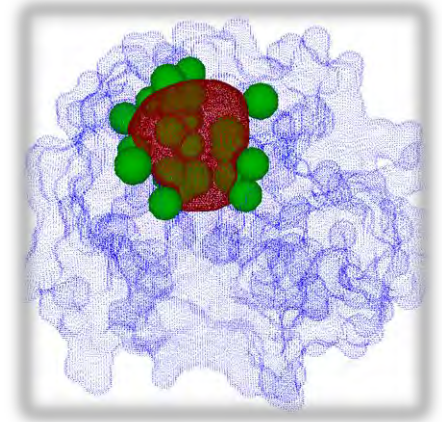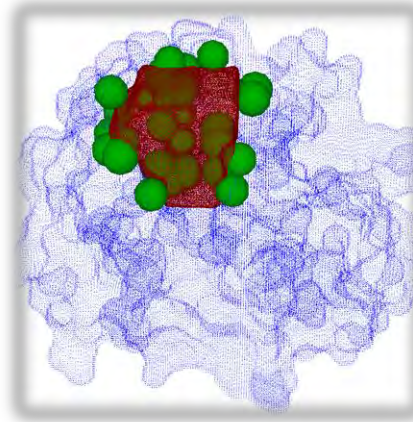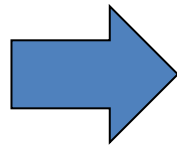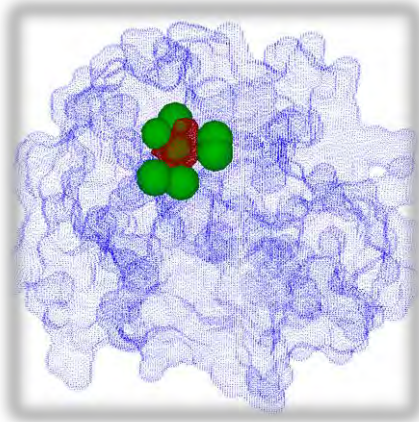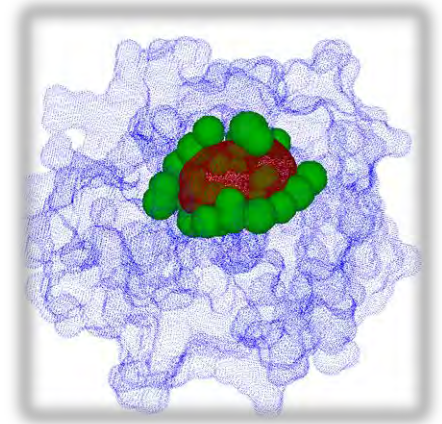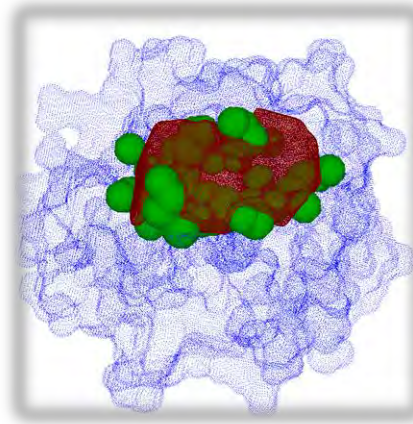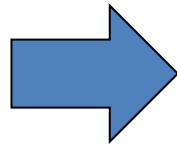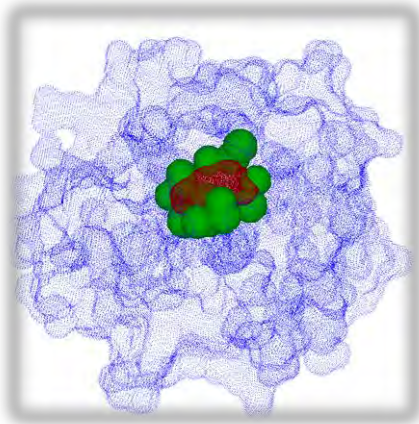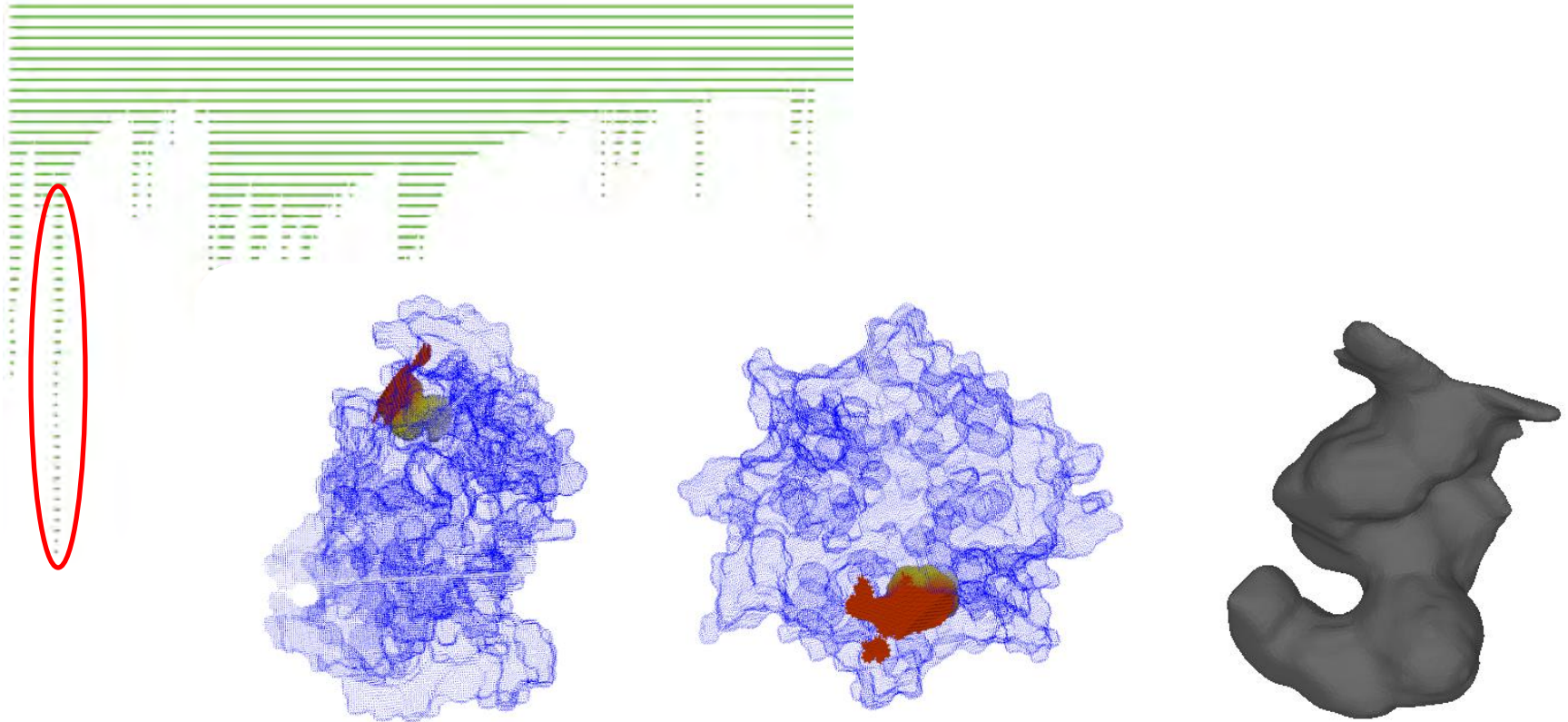
The main pocket

The second main pocket
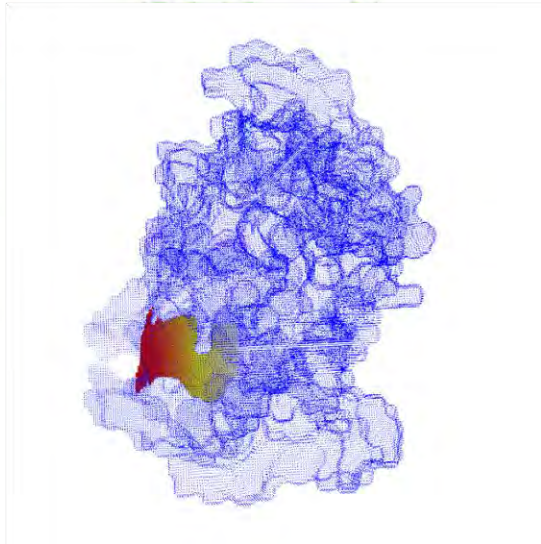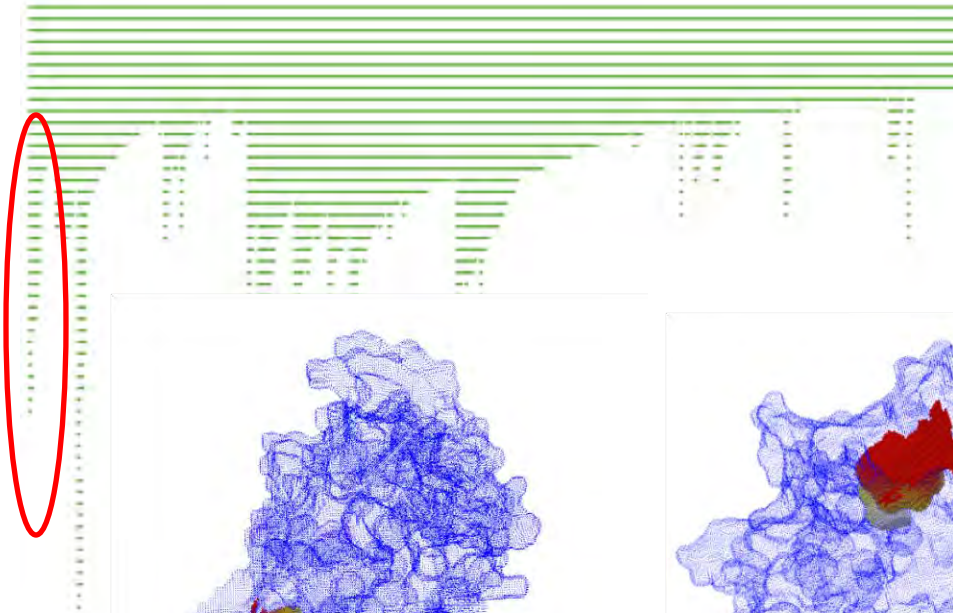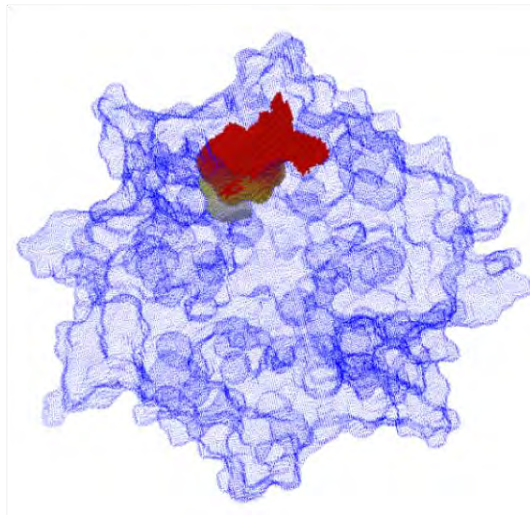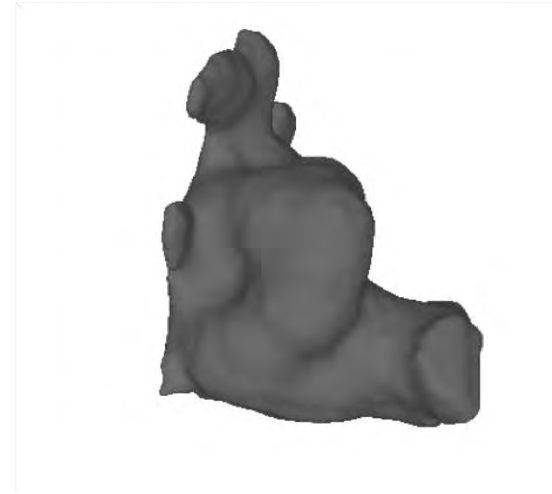
The main pocket of 1MK5
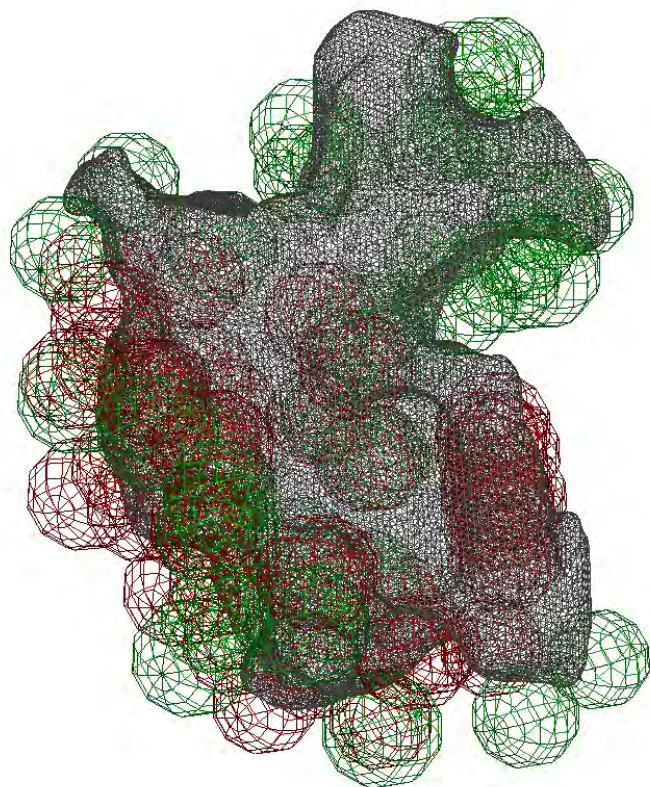
- Adjacent atoms
  - green from $CAST_p$
  - Red and green from PPS

1MK5
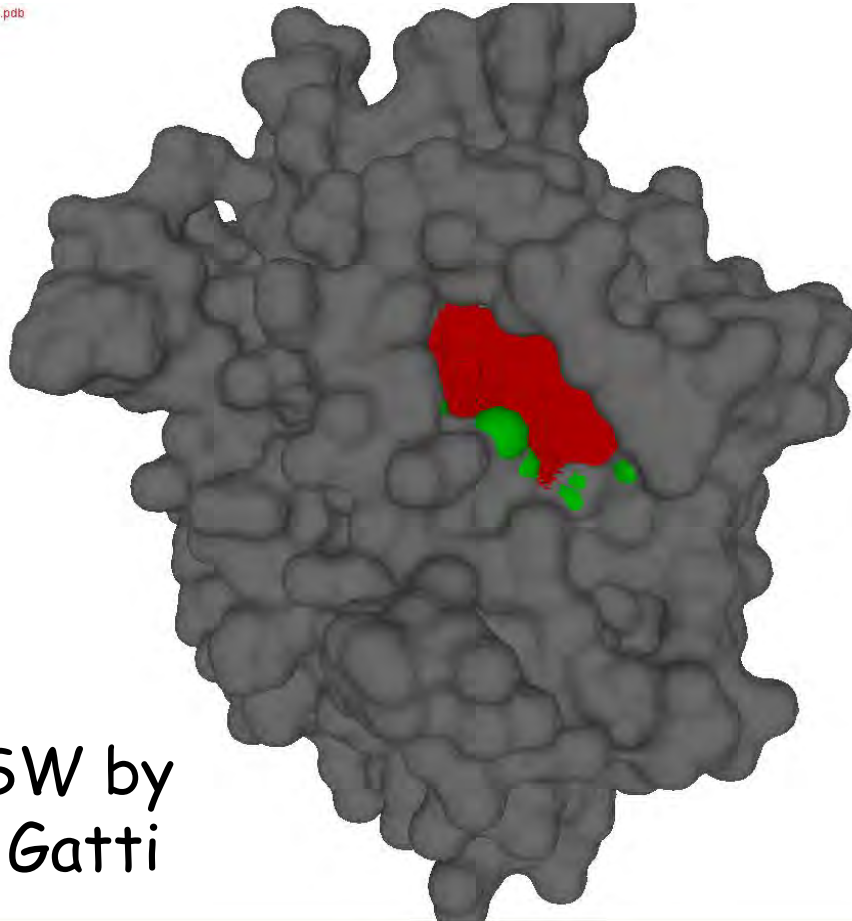
Credits: SW by
Riccardo Gatti

Protein Inspector – finding protrusions

Segmentation of protrusions

for a 2D shape
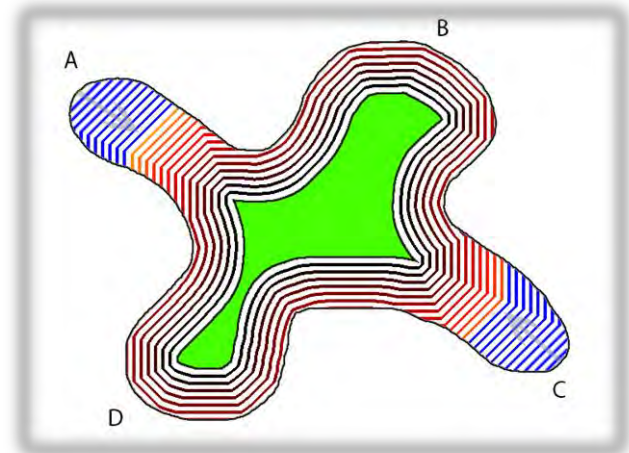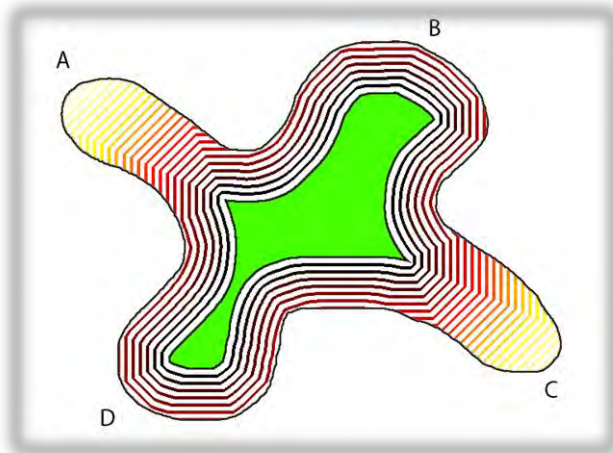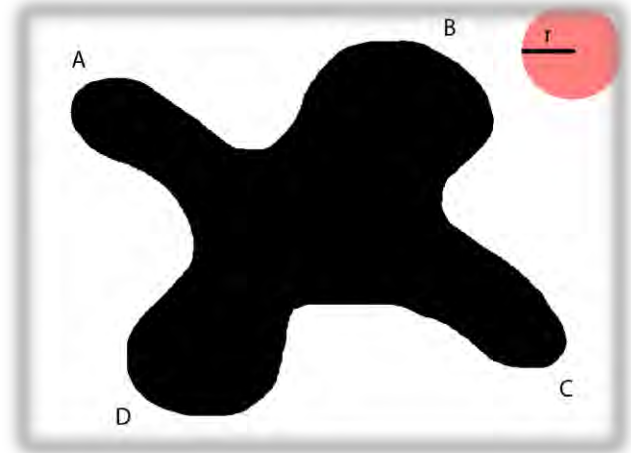
Protein Inspector – finding protrusions
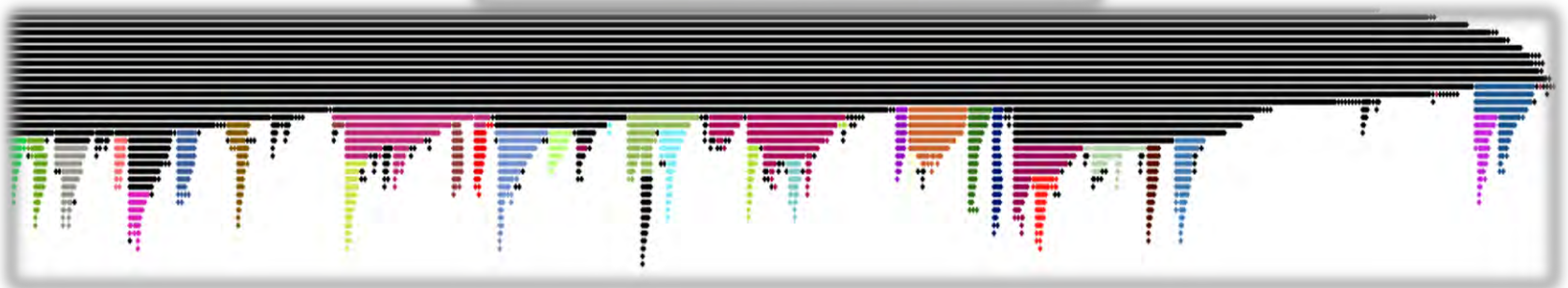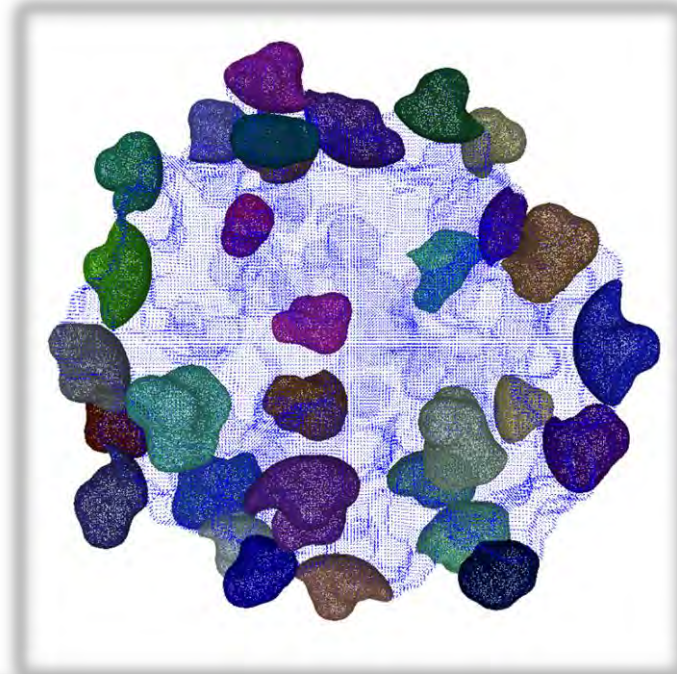
Protein inspector - user interface
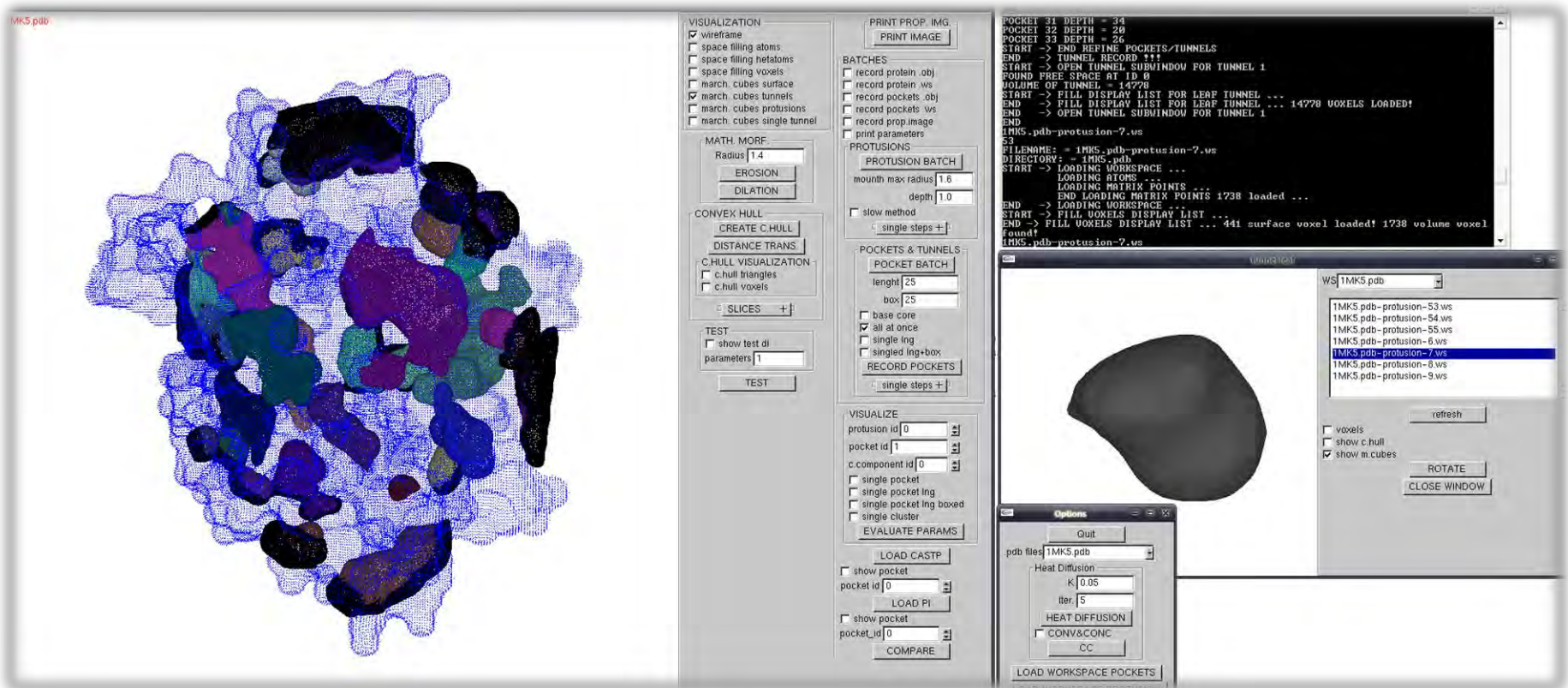
2D sketch for searching protuberances
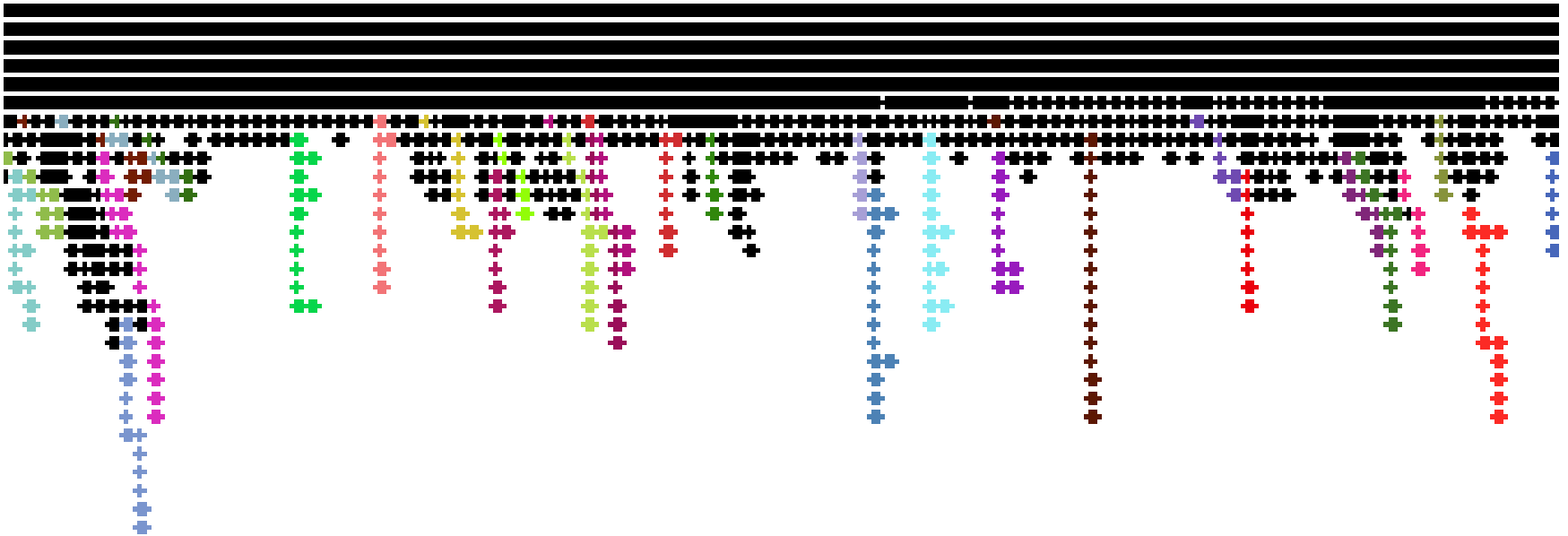
Active sites matching

- Two data structures are proposed:

  - the Concavity Tree

    Arcelli C, Sanniti di Baja G (1978) "Polygonal covering and concavity tree of binary digital pictures", Proceeding International Conference MECO '78, Athens, pp. 292–297.

  - the Extended Gaussian Image

    Hu, Z., · Chung, R., · Fung K. S. M., (2010), "EC-EGI: enriched complex EGI for 3D shape registration", Machine Vision and Applications, 2, 177–188.

Concavity Tree

Concavity Tree

First level

Second level

Third level

Fourth level
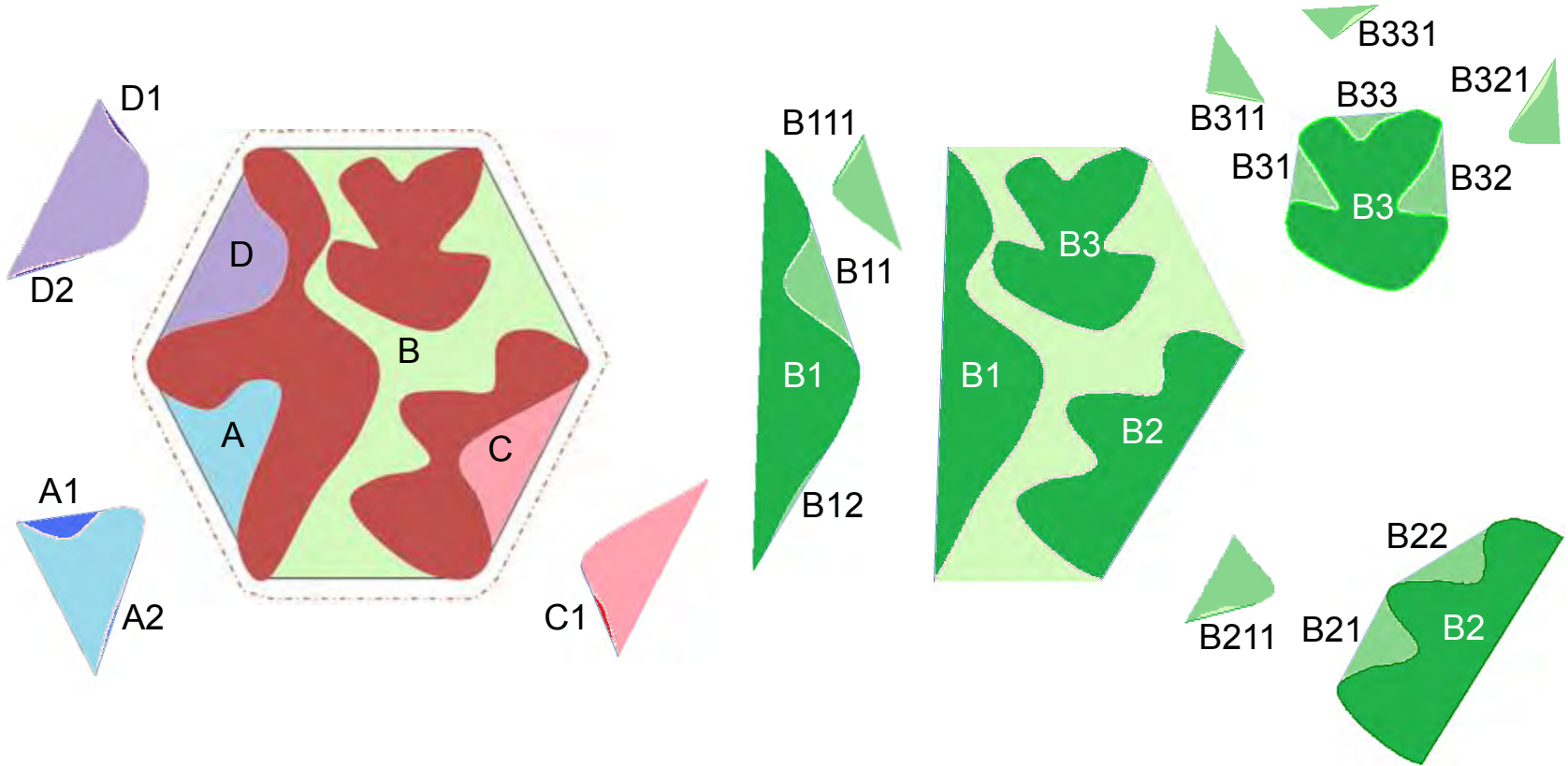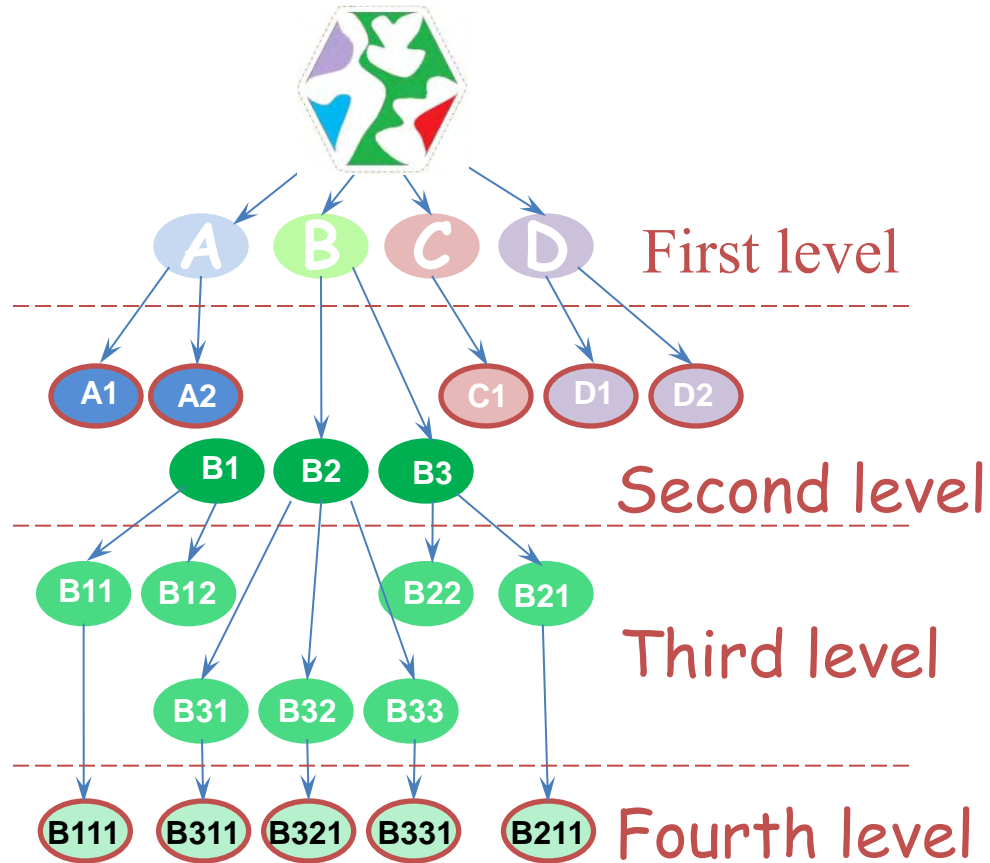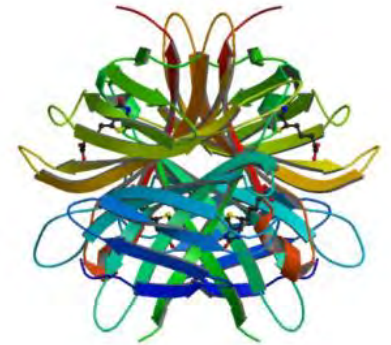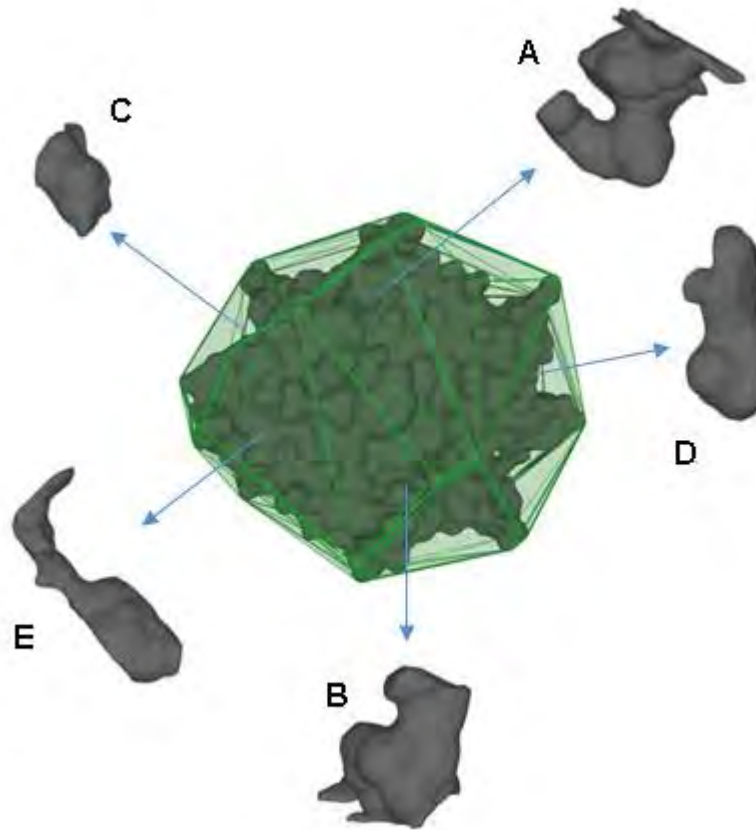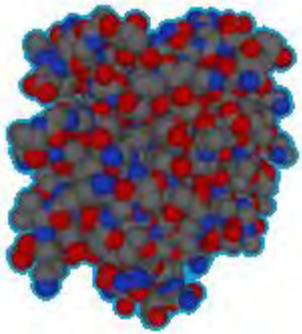
Protein 1MK5

Protein 1MK5 concavity tree
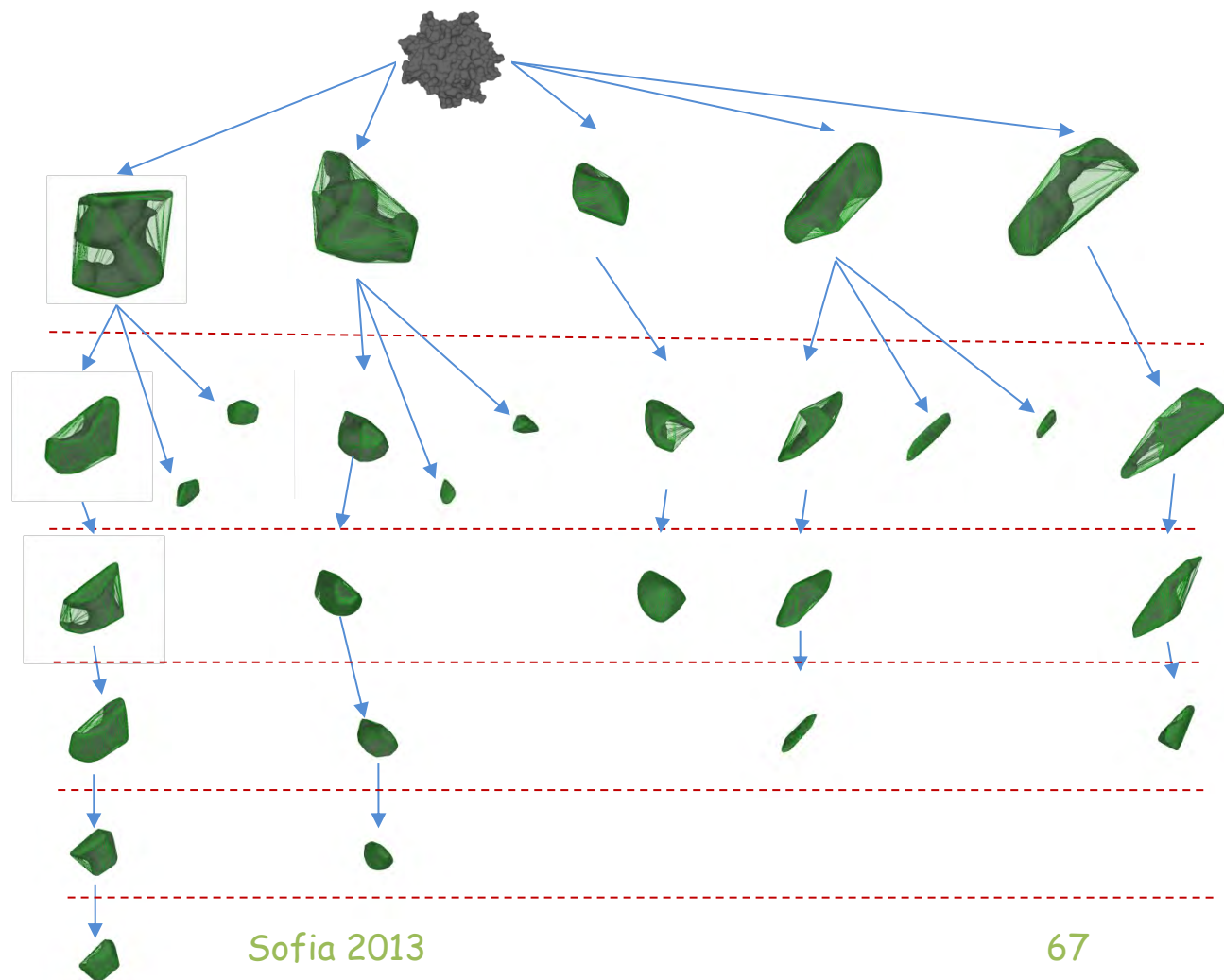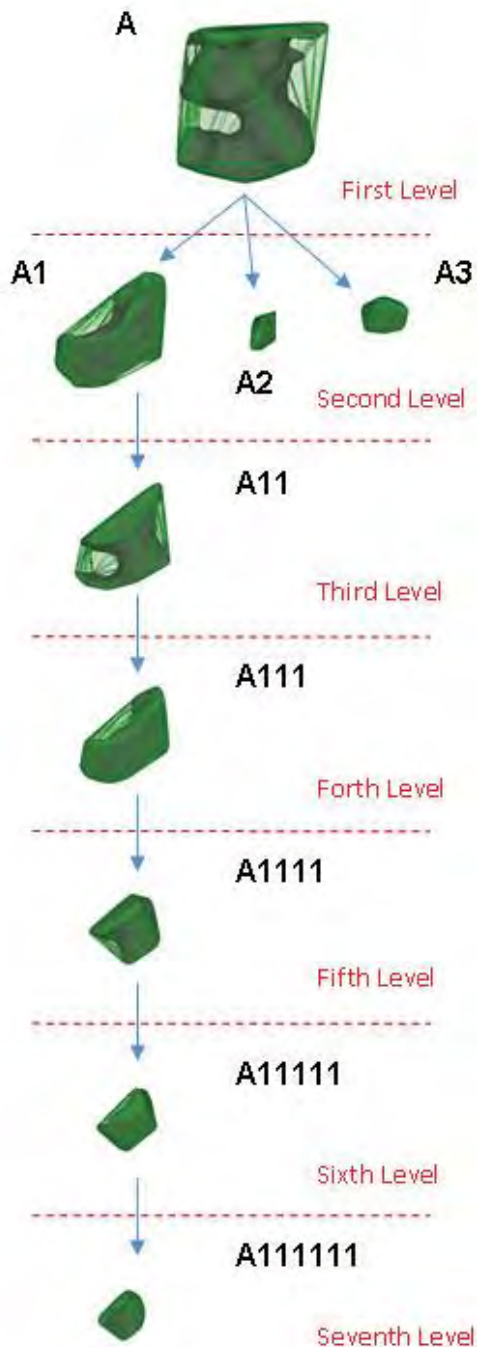
Concavity tree
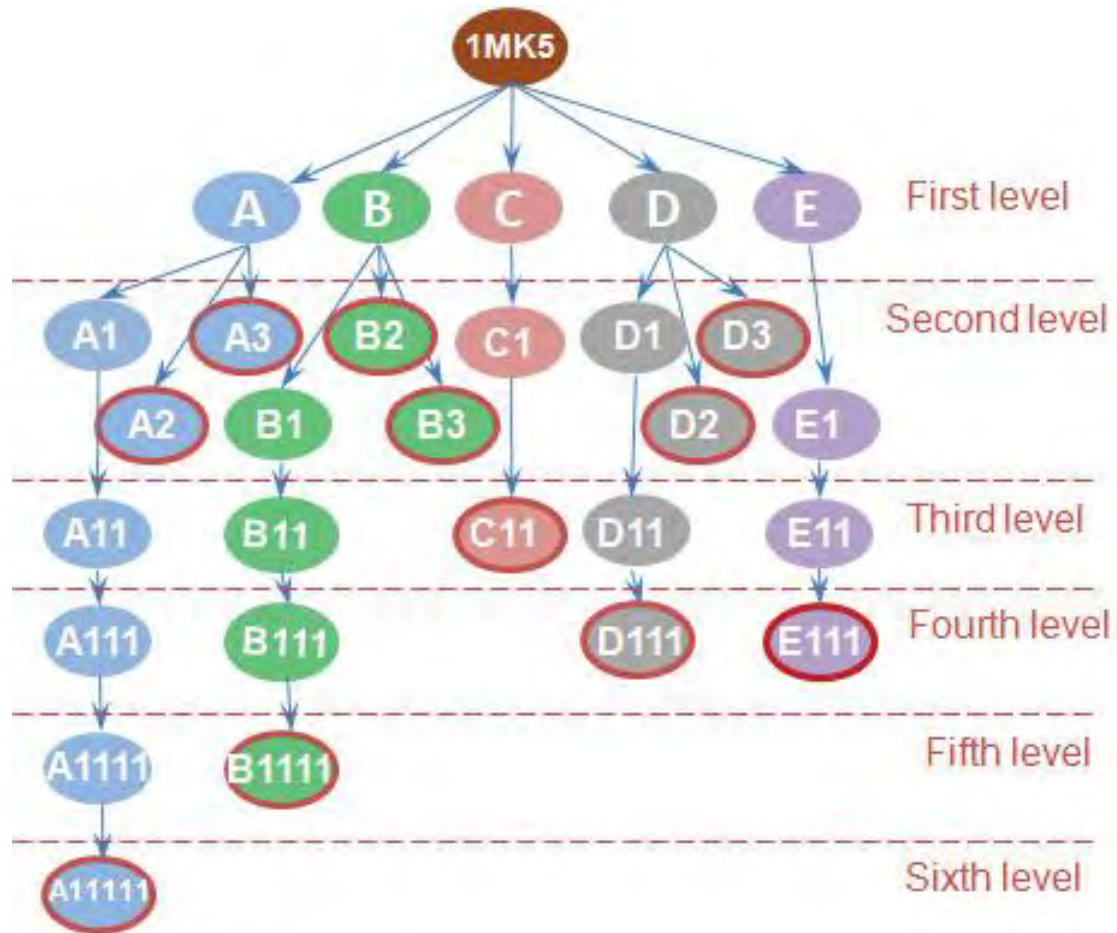
CT: node content

- *Volume*
- *Surface to Volume Ratio*
- *Skewness of Height Distribution*
- *Kurtosis of Height Distribution*
- *Mouth Aperture*

- *Travel Depth*
- *Top Five Peaks and Valleys*
- *Summit Density*
- *Mean Summit Curvatures*
- *Interfacial Area Ratio*
- *Residue Conservation*

Extended Gaussian Image

The EGI of a 3D object or shape is an orientation histogram that records the distribution of surface area with respect to surface orientation.

Gauss mapping

(a) Cube

(b) EGI of cube
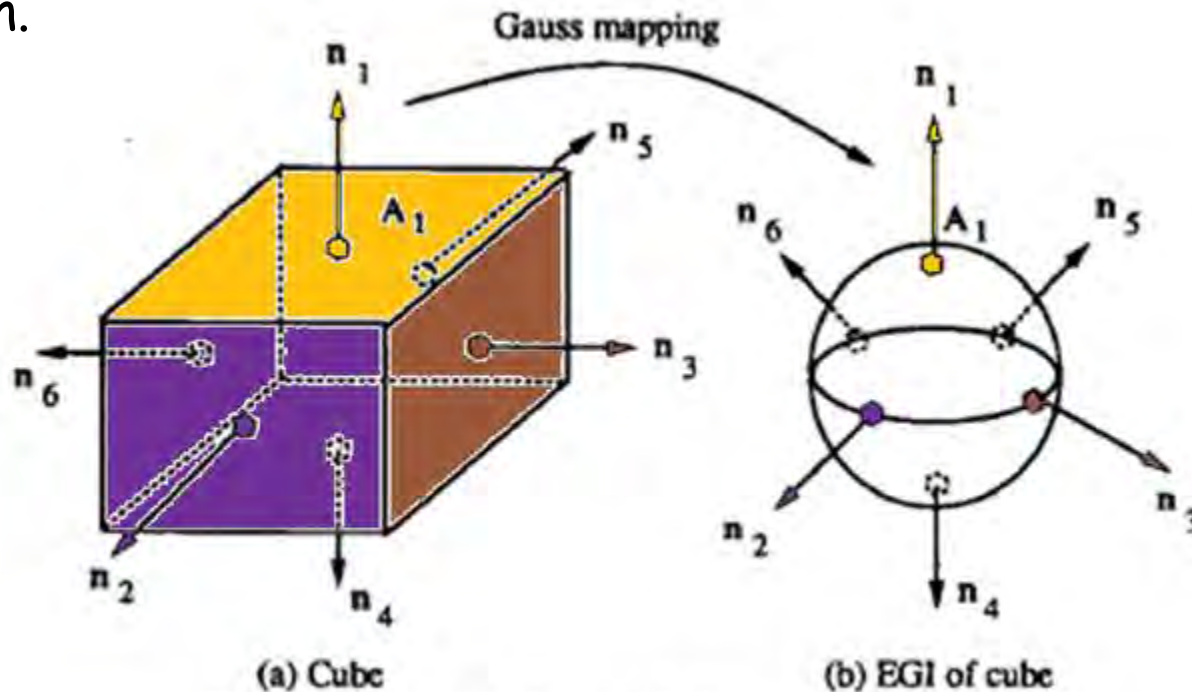
(Note: The weight is shown only for normal n₁ for clarity)

B.K.P. Horn, Extended Gaussian Images, 1984

Extended Gaussian Image

Hierarchical Search
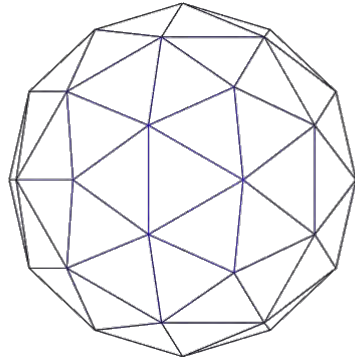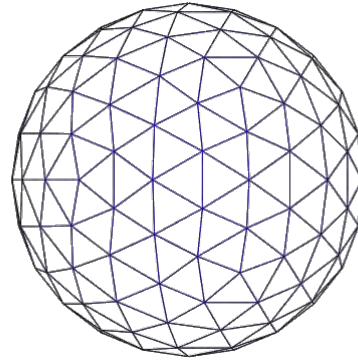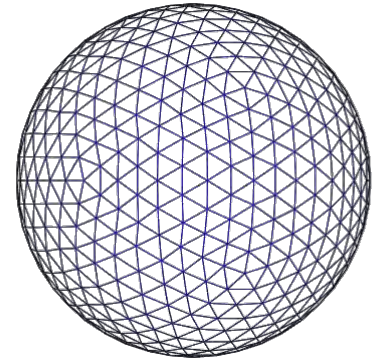
20 Faces	80 Faces	320 Faces	1280 Faces
Icosahedron			

Funk et al., SIGGRAPH 2004

Hierarchical Search

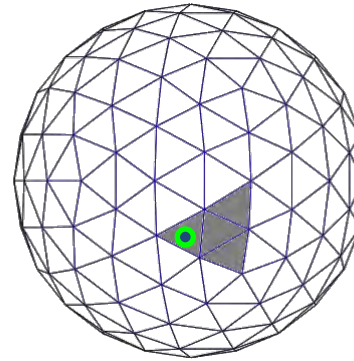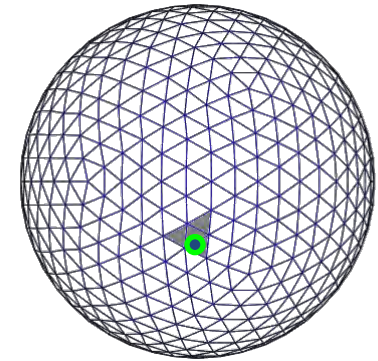| 20 Faces | 80 Faces | 320 Faces | 1280 Faces |
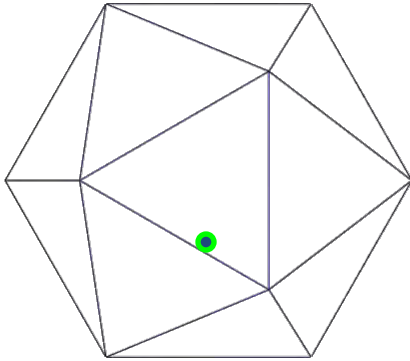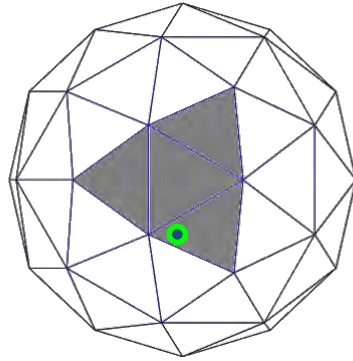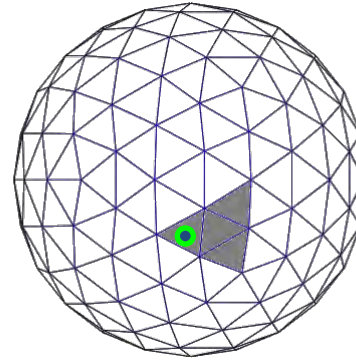
Icosahedron
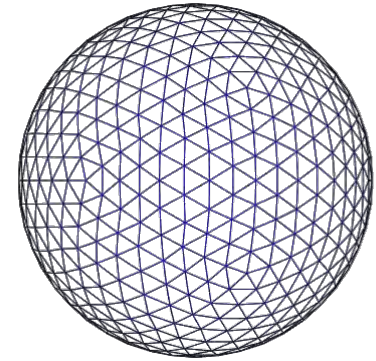
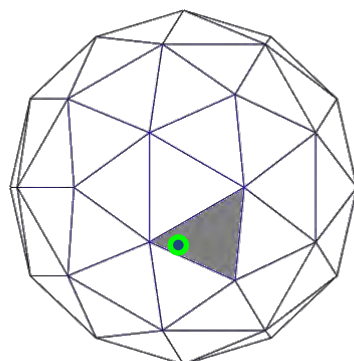*Funk et al., SIGGRAPH 2004*
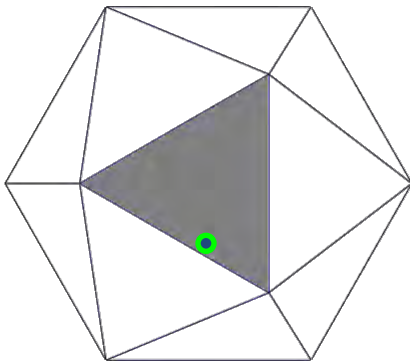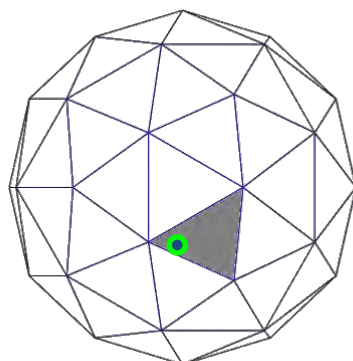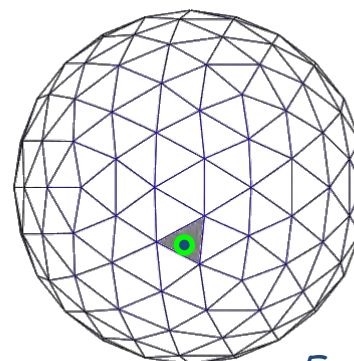
Hierarchical Search

20 Faces
Icosahedron

80 Faces
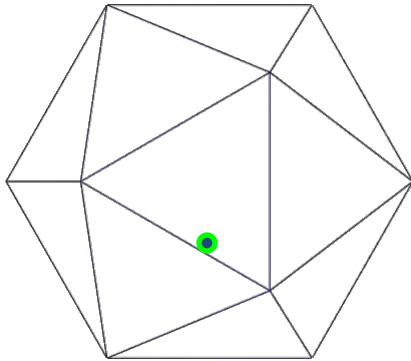
320 Faces

1280 Faces

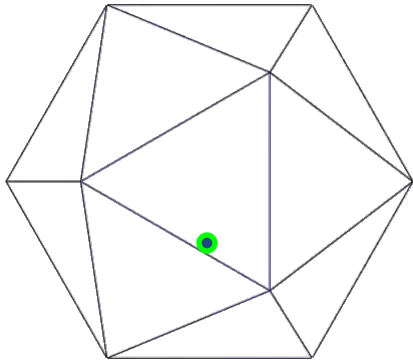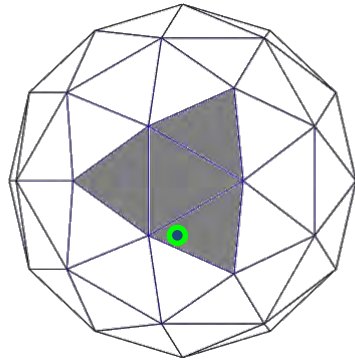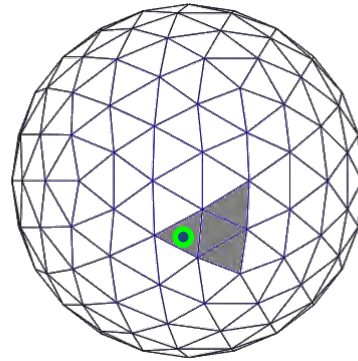*Funk et al., SIGGRAPH 2004*

Hierarchical Search

20 Faces
Icosahedron

80 Faces

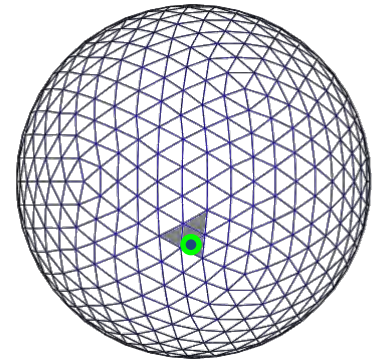320 Faces

1280 Faces

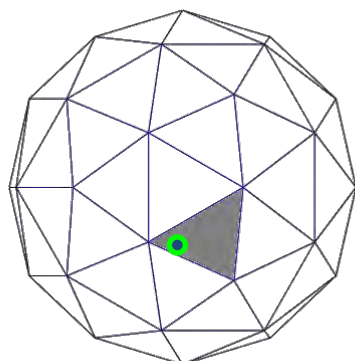*Funk et al., SIGGRAPH 2004*

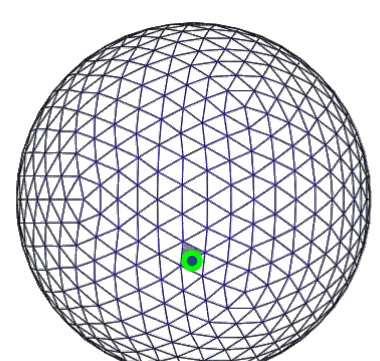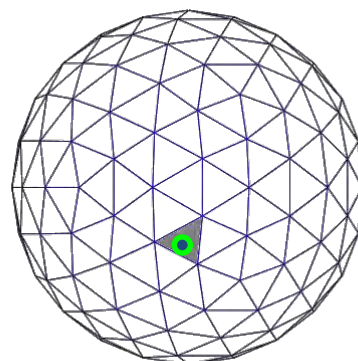# Hierarchical Search

20 Faces
Icosahedron

80 Faces

320 Faces

1280 Faces
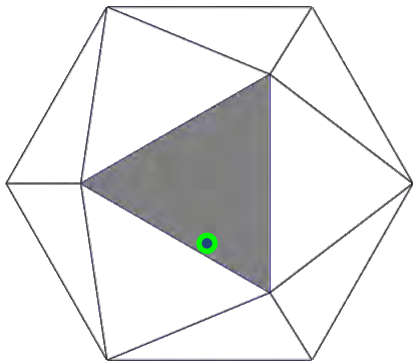
Extended Gaussian Image

- EGI The EGI of a 3D object or shape is an orientation histogram that records the distribution of surface area with respect to surface orientation.

Horn, B.K.P.: Extended Gaussian images. In: Proceedings of the IEEE 72, 1671–1686 (1984).

- CEGI The Complex EGI encodes each surface patch's signed perpendicular distance from the reference coordinate center.

Kang, S.B., Ikeuchi, K.: Determining 3-D object pose using the complex extended Gaussian image. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 580–585 (1991).

- MEGI The More Extended Gaussian Image (MEGI) model consists of a set of position vectors for surfaces and their normal vectors.

H. Matsuo and A. Iwata, "3-D Object Recognition Using MEGI Model from Range Data." Proc. 12th Int'l Conf. Pattern Recognition, pp. 843-846, Jerusalem, Israel, Oct. 1994.

- MSEGI The multi shell EGI

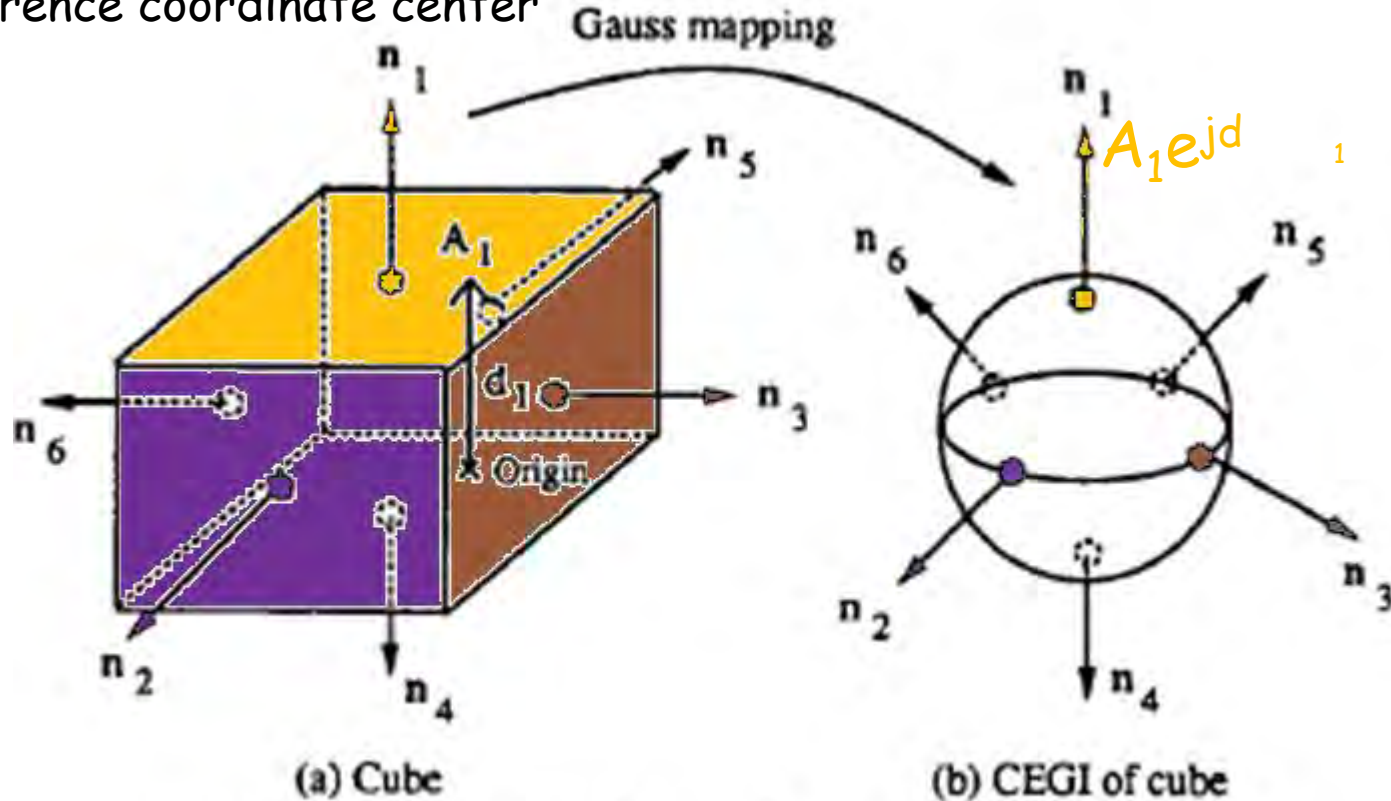Dingwen Wang[1,2], Jiqi Zhang[2], Hau-San Wong[2], and Yuanxiang Li, 3D Model Retrieval Based on Multi-Shell Extended Gaussian Image, G. Qiu et al. (Eds.): VISUAL 2007, LNCS 4781, pp. 426–437, 2007.

- ECEGI The Enriched CEGI encodes each surface patch's signed with its 3D position.

Zhaozheng Hu · Ronald Chung · Kenneth S. M. Fung: EC-EGI: enriched complex EGI for 3D shape registration, Machine Vision and Applications (2010) 21:177–188.
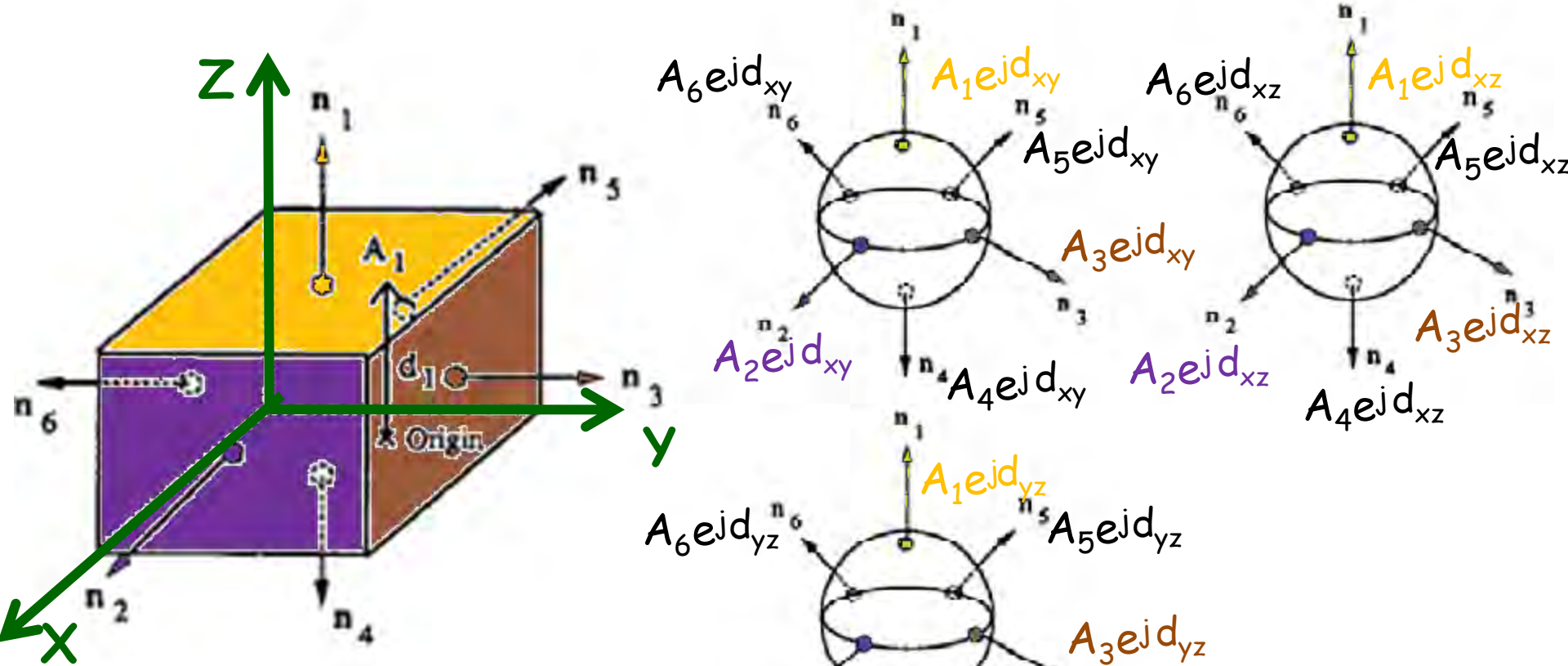
Complex EGI (C-GEI)

CEGI encodes each surface patch's signed perpendicular distance from the reference coordinate center

Gauss mapping

$A_1 e^{j d_1}$

n_1 n_5 A_1 d_1 n_3 Origin n_6 n_2 n_4

(a) Cube

n_1 n_6 n_5 n_2 n_3 n_4

(b) CEGI of cube

(Note: The weight is shown only for normal n_1 for clarity)

Sing Bing Kang and Katsushi Ikeuchi, 3-D Object Pose Determination Using Complex EGI, 1993
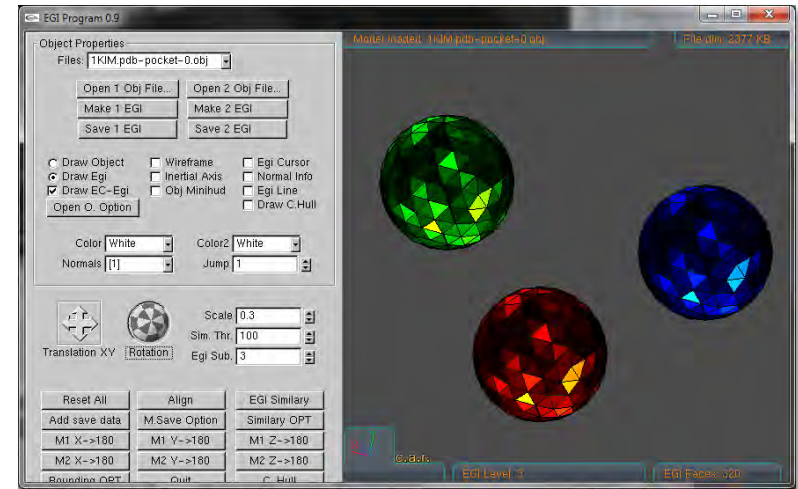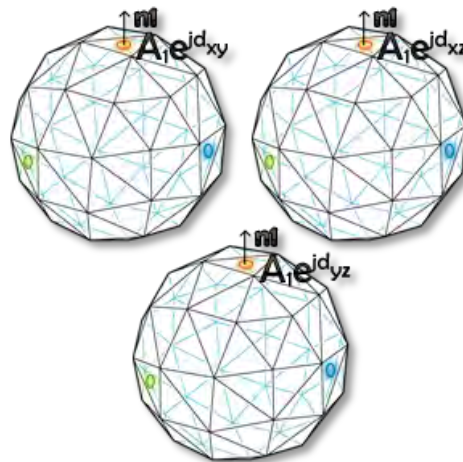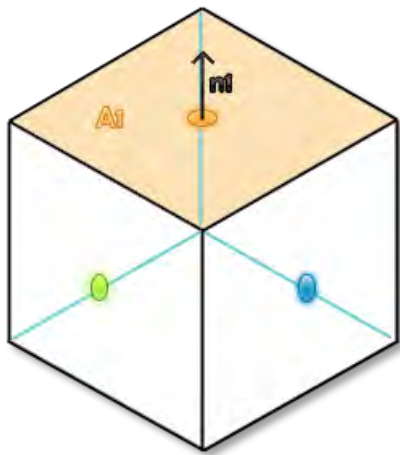
Enriched C-EGI (EC-EGI)

The EC-EGI encodes each surface patch's signed with its 3D position

Hu, Z., · Chung, R., · Fung K. S. M., (2010), "EC-EGI: enriched complex EGI for 3D shape registration", Machine Vision and Applications, 2, 177–188.
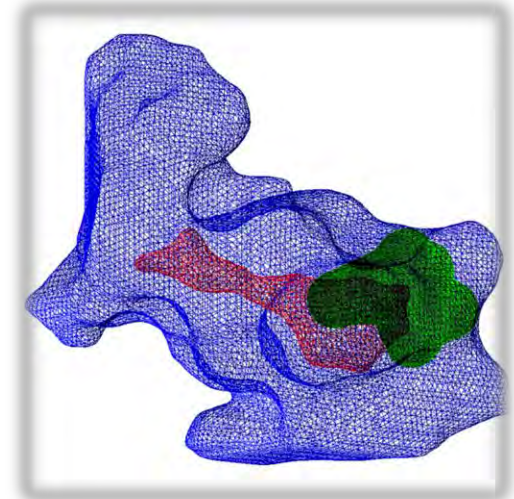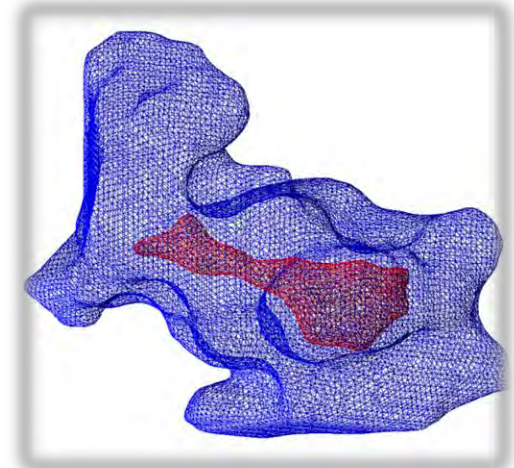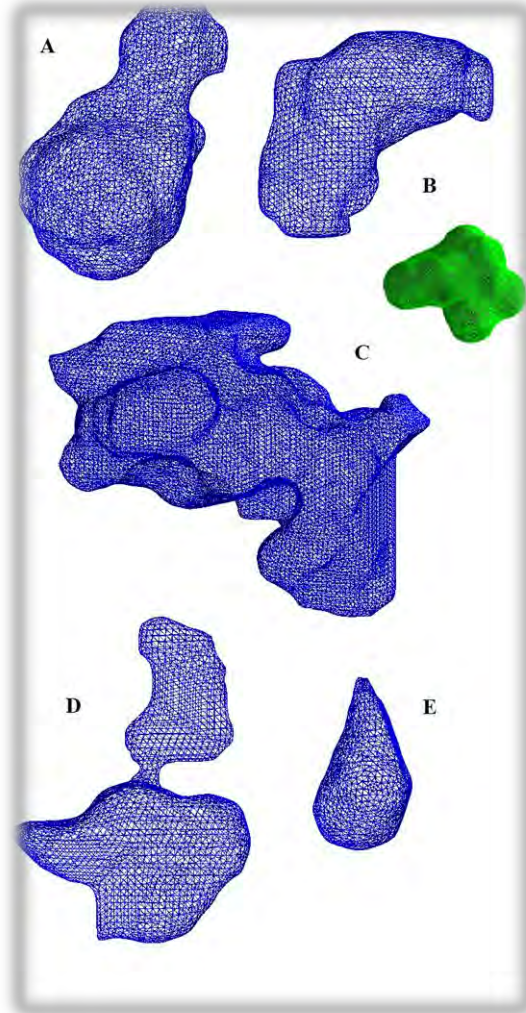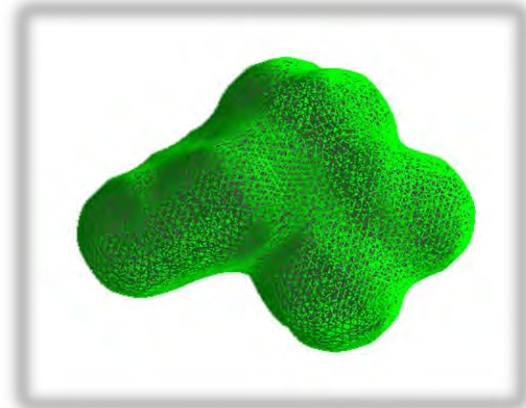
Enriched C-E.G.I.

EGI in progress

Credits: SW by
Alessandro Gaggia

Protein-Ligand Interaction
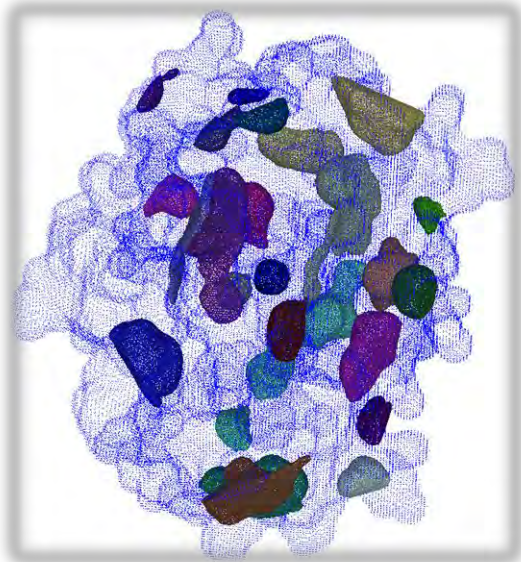
CAPRI: Critical Assessment of PRediction of Interactions

- CAPRI (Critical Assessment of PRedicted Interaction) is an international effort, aimed at objectively assessing the performance of these methods by inviting developers to test their algorithms on the same protein targets and objectively evaluating the results.

Capri