Distributed computations in physical bioinformatics tasks

Akishina T.P., Zrelov P.V., Ivanov V.V. (JINR, Dubna),
Polozov R.V. (Institute of Theoretical and Experimental Biophysics RAS, Pushchino),
Sivozhelezov V.S. (Institute of Cell Biophysics RAS, Pushchino; Chair of Biophysics, University of Genova, Italy)

Why is GRID adequate for biology

- GRID is well suited for multiple similar computation tasks
- Life is about interactions of biomolecules; *interactome* is the whole set of molecular interactions in cells
- How many? In terms of proteins... Let’s assume each gene encodes one protein... Circa 35,000 genes...combinations (35,000; 2).... about 5 million.
- Multiple (up to millions) analogous computational tasks; requirements for each of them easy to evaluate sometimes from input file size.
- The above is for *all* proteins in *one* organism; Similar estimates result for comparing *same* protein (and its interactions) in *different* species, because there are from 2 mln to 100 mln biological species [Society For Conservation Biology (2003, May 26). Just How Many Species Are There, Anyway?. ScienceDaily. Retrieved June 29, 2008, from http://www.sciencedaily.com- /releases/2003/05/030526103731.htm]

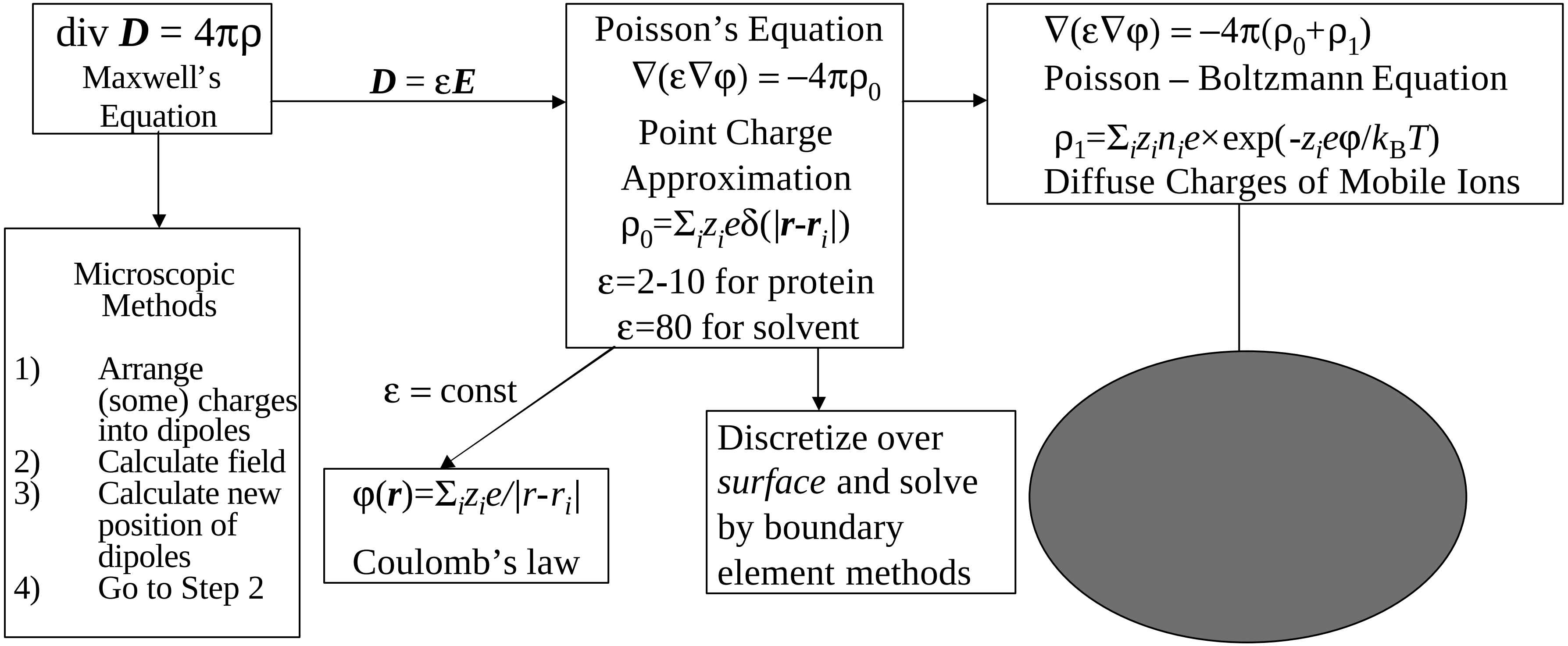
How do molecules recognize each other ?

- Earliest stage of bi(o)molecular recognition is ***Coulombic electrostatic*** interactions, because Coulombic forces decay with distance slower than other forces operating in biological cells
- Distances over 5 Å, where only electrostatic forces contribute
- With Coulombic electrostatic calculations alone, at least, we can say protein X *will not* interact with protein Y; that's <u>already important</u>, considering that the experimental procedures used in proteomics are prone to false positives.

ELECTROSTATICS
Physical Grounds For Biological Significance

- Electrostatic forces **decay slowly** with distance
- Electric charge and potential have **signs** responsible for interaction specificity
- The **Irnshaw theorem**: any system of charges is unstable – so there is a **driving force**.
- The **Free Energy theorem**: Electrostatic field energy contributes additively to **free energy**.
- **Superposition principle** facilitates understanding effects of mutations

Electrostatics - Hierarchy



$\mathrm{div}\,\boldsymbol{D} = 4\pi\rho$

Maxwell's Equation

$\boldsymbol{D} = \varepsilon \boldsymbol{E}$

Poisson's Equation

$\nabla(\varepsilon\nabla\varphi) = -4\pi\rho_0$

Point Charge Approximation

$\rho_0=\Sigma_i z_i e\delta(|\boldsymbol{r}\text{-}\boldsymbol{r}_i|)$

ε=2-10 for protein

ε=80 for solvent

$\nabla(\varepsilon\nabla\varphi) = -4\pi(\rho_0+\rho_1)$

Poisson – Boltzmann Equation

$\rho_1=\Sigma_i z_i n_i e\times\exp(-z_i e\varphi/k_B T)$

Diffuse Charges of Mobile Ions

Microscopic Methods

1) Arrange (some) charges into dipoles
2) Calculate field
3) Calculate new position of dipoles
4) Go to Step 2

ε = const

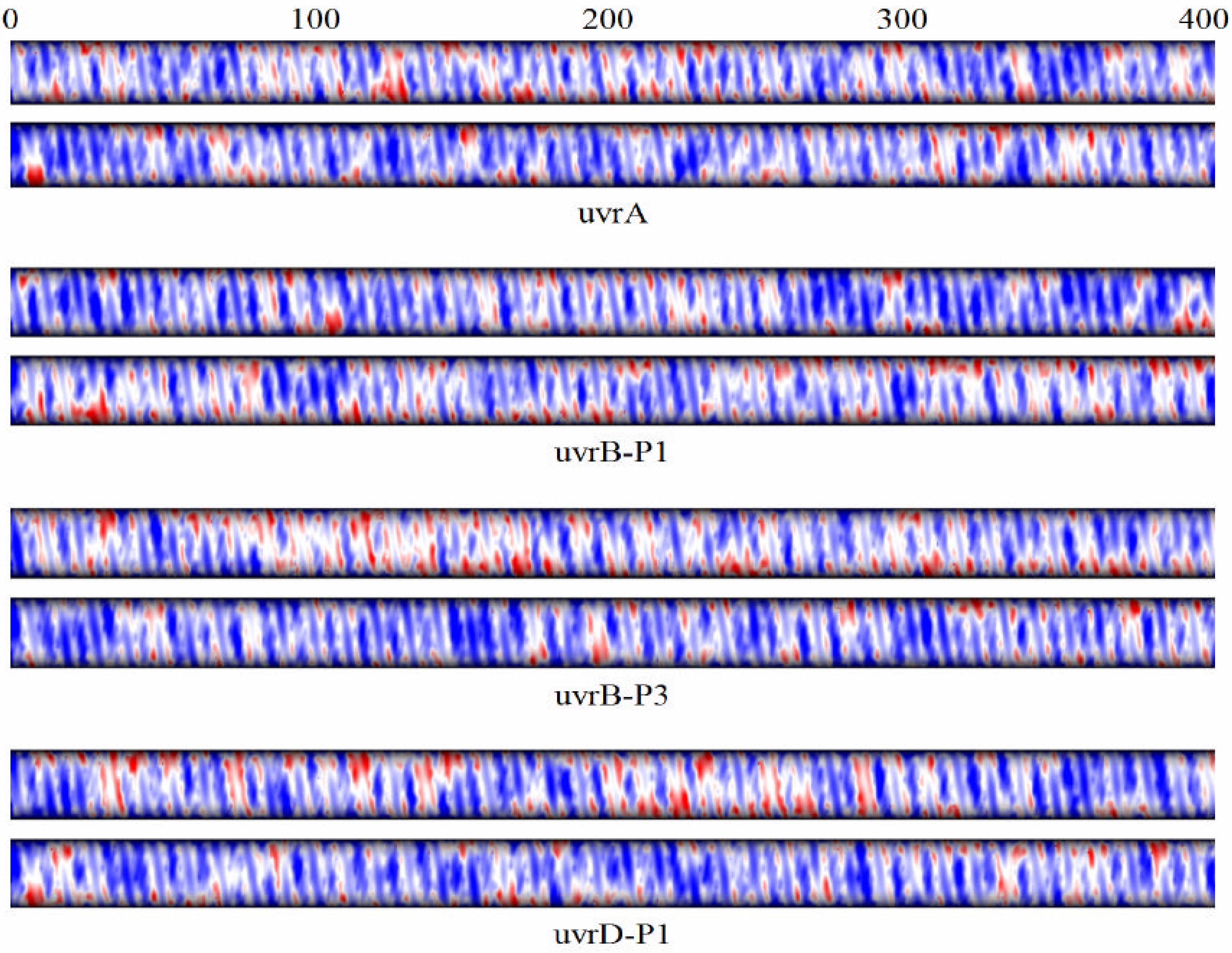
$\varphi(\boldsymbol{r})=\Sigma_i z_i e/|r\text{-}r_i|$

Coulomb's law

Discretize over *surface* and solve by boundary element methods

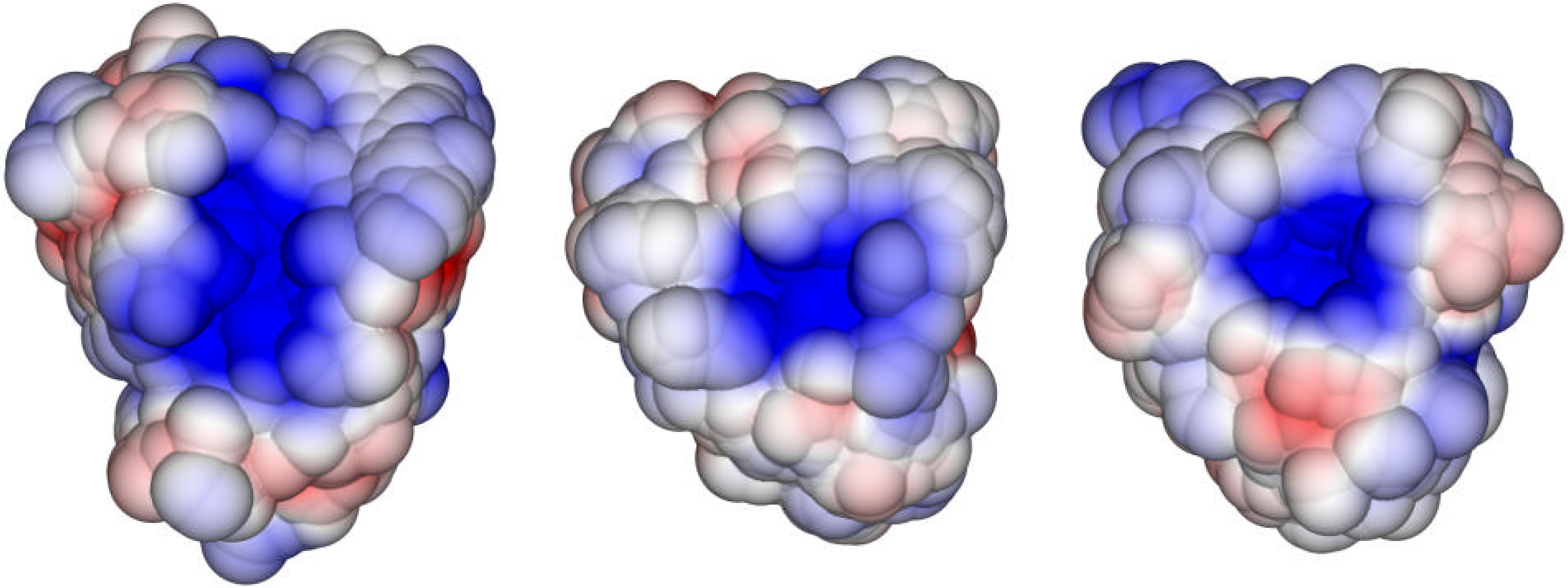
Genomic DNA

- Main control of protein biosynthesis is at the stage of transcription
- Transcription is controlled by so called ***promoter regions***, typically one promoter region for one gene. Regulation proceeds via proteins, the so-called ***transcription factors***, which bind to promoter DNA thus blocking (or unblocking) the access for the the catalytic machinery that “reads” each gene.
- Promoter regions are several hundreds of DNA base pairs long
- Need to look at all promoters for the given biological species
- Electrostatic potential is mapped onto a cylinder encompassing the DNA molecule; red color denotes higher probability of capturing the transcription factors.

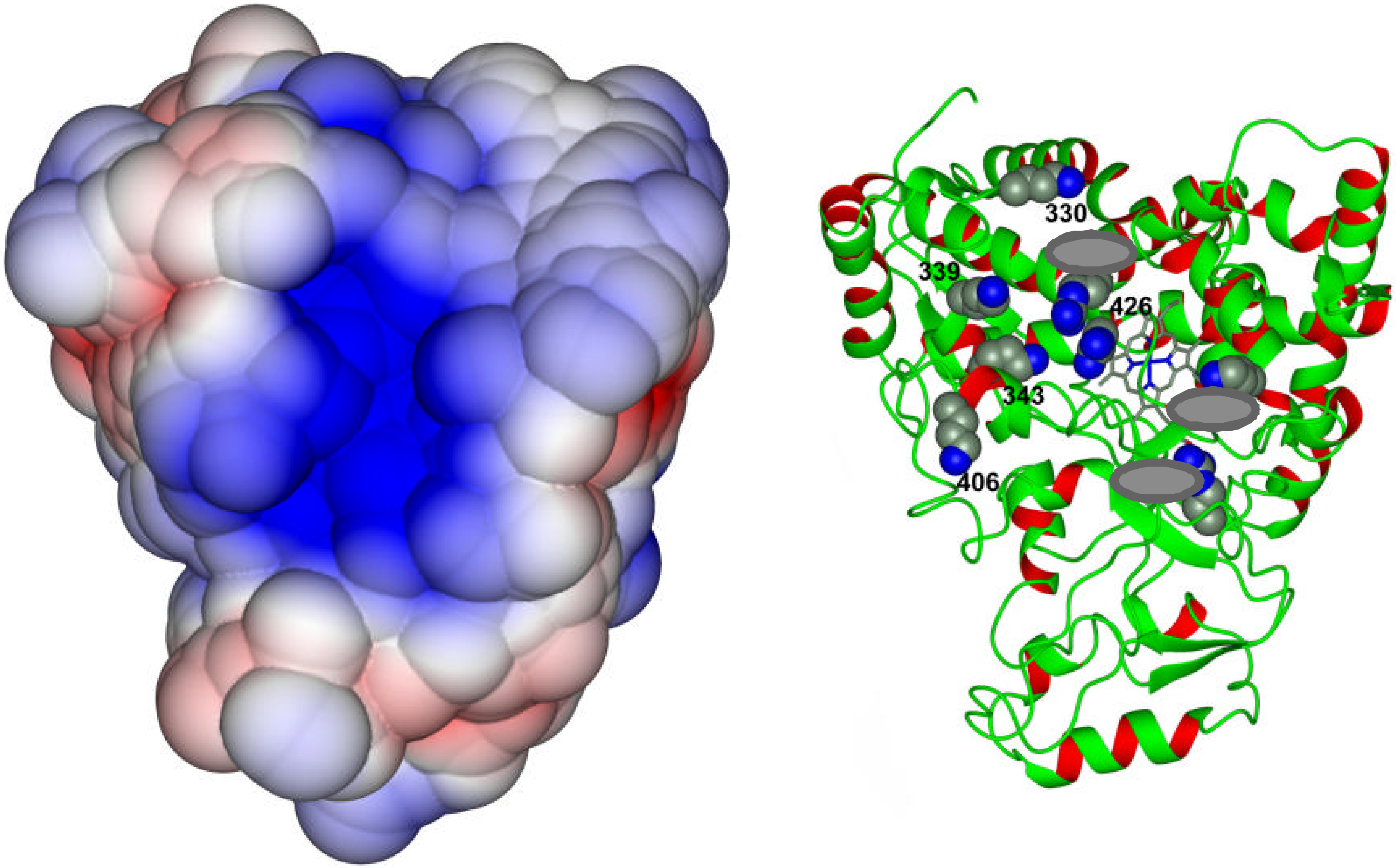

0
100
200
300
400
uvrA
uvrB-P1
uvrB-P3
uvrD-P1

Same protein in different organisms - Cytochrome P450

- Metabolize organic substances by inserting a single oxygen atoms into the substrate
- Need extra electron (s) for function
- **Electrons are delivered by the special protein cofactor**
- **Cofactors are negatively charged and ~100 aa long i.e. approximate radius of cofactor is 10-15 Å**
- For P450scc, only a homology model is available using a very remote homolog, P450cam.
- Positive potential patches around P450scc
- Contributing amino acids
- Check if mutation of the identified amino acid affect the cofactor binding and the overall rate.



Electrostatic potentials around Cytochromes P450 and its two VAST neighbors closest by alignment length. Left to right: P450SCC, P450BM3, P4502B4



Cytochrome P450scc: electrostatic potential (left), contributing residues on a ribbon diagram (right). Marked are residues aligned in the amino acid sequences of Cytochromes P450.

Comparison of proteins according to their electrostatic potentials

- Conserved electric potentials do not imply conserved charged amino acid residues
- Comparison of proteins by amino acid sequence is insufficient (recognized in 1990s) for understanding function and evolution
- Comparison of 3D structures **is not sufficient either**. To understand protein function, we need to compare **physical properties**.
- Hence the need for a new interdisciplinary area – ***Physical Bioinformatics*** or ***Biophysical Informatics***.

Nanotech application: biosensor development

- Molecular modeling for nanotechnology purposes = computational nanotechnology
- Why is it necessary for biosensor development ?
- Why is GRID relevant ?
- Why electrostatic (and physical bioinfo in general) component of computational biotechnology important for biosensors?

What does biosensor development look like *now* ?

- Biosensor development is about immobilization of enzymes
- Biosensors are developed in the same manner as 30-40 years ago, mostly by trial and error
- Ignored is the huge body of data on structure and function of enzymes (special kind of proteins)
- Ignored is another huge body of data on immobilization of enzymes, if it is not directly relevant for the chosen enzyme and the chosen method of immobilization

Biosensor development

How it *is* done

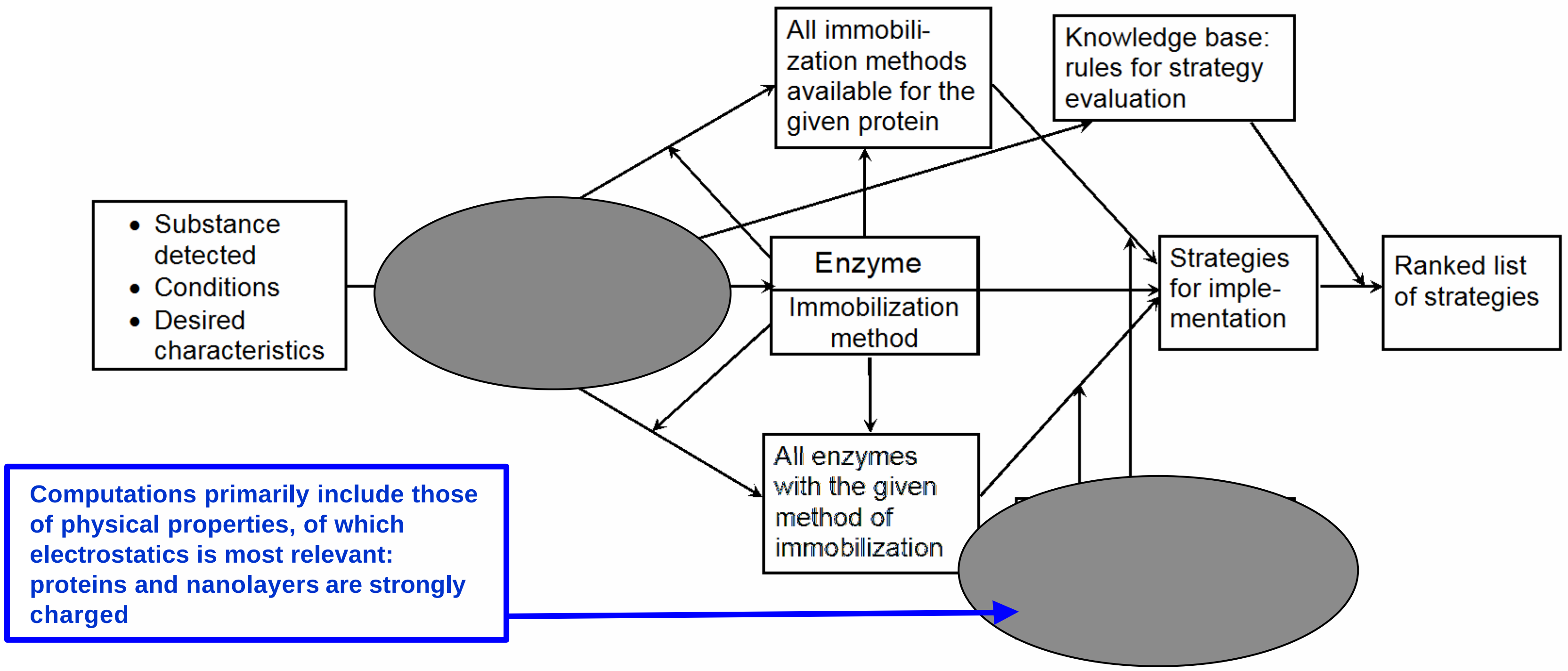
- Question asked: what biosensor can I make with the given ***enzyme*** and method of immobilization
- Arbitrary choice of enzyme and method of immobilization

How it ***should be*** done

- Question asked: what biosensor can I make for the given ***target substance*** and the given requirements?
- Explore all relevant enzymes and all methods of immobilization available for those enzymes

Conclusion: database of immobilized enzymes is necessary for biosensor development

... and not only a database, but an expert system!


• Substance detected
• Conditions
• Desired characteristics
All immobili-zation methods available for the given protein
Knowledge base: rules for strategy evaluation
Enzyme
Immobilization method
Strategies for imple-mentation
Ranked list of strategies
All enzymes with the given method of immobilization

Computations primarily include those of physical properties, of which electrostatics is most relevant: proteins and nanolayers are strongly charged

Flow chart of the expert system for biosensor development.

- Each strategy for biosensor implementation includes: the enzyme, the immobilization method, and design of the biosensor
- Rules for evaluation of the strategies are contained in the knowledge base.)
- Computations are necessary to evaluate feasibility of the biosensor, when the database contains the data on the ***given enzyme, but different immobilization method from the given one***, or vice versa.
- Then, solution of hundreds to thousands similar problems is required (GRID).

Tasks for GRID in physical bioinformatics

- ribosomal protein biosynthesis, where subnanometer-precision calculation of the electrostatic field of the ribosome are necessary to model the process of amino acid delivery to the active site of protein biosynthesis, as well as to account for the influence of electrostatic field of the ribosome to the folding of the nascent polypeptide chain into the functional (native) protein structure

- calculations of physicochemical properties such as electrostatic potentials, distributions of hydrophobic regions, etc., of long (thousands basepairs) of the regulatory fragments of genomic DNA, especially promoters/operators, signal sequences, etc., and recognition of those DNA fragments by proteins

- comparative analysis and classification of large protein families according to their physicochemical properties and structures,

- calculation of physicochemical properties of the collagen molecule and modeling of collagen packing into fibrils and higher-level structures

- calculation of physicochemical properties of the beta layers of amyloid peptide molecules and modeling of formation of senile plaques and transmembrane channels by amyloid peptides.