Biological Protein-protein Interaction Prediction using Binding Free Energies and Linear Dimensionality Reduction

Luis Rueda
School of Computer Science
University of Windsor
Canada

* Joint work with Sridip Banerjee and Md. Mominul Aziz , University of Windsor and Carolina Gárate , University of Concepción, Chile

Protein-protein Interaction

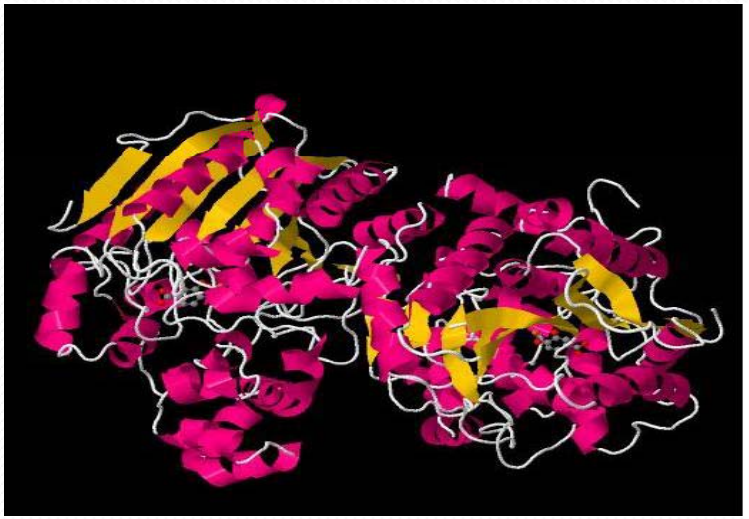
- Important roles in many biological processes.
  - signal transduction
  - to form a protein complex
  - protein may carry another protein
  - protein may interact to modify another protein.

To know biological process,

understanding of diseases,

for the sake of treatment,

drug development, etc.

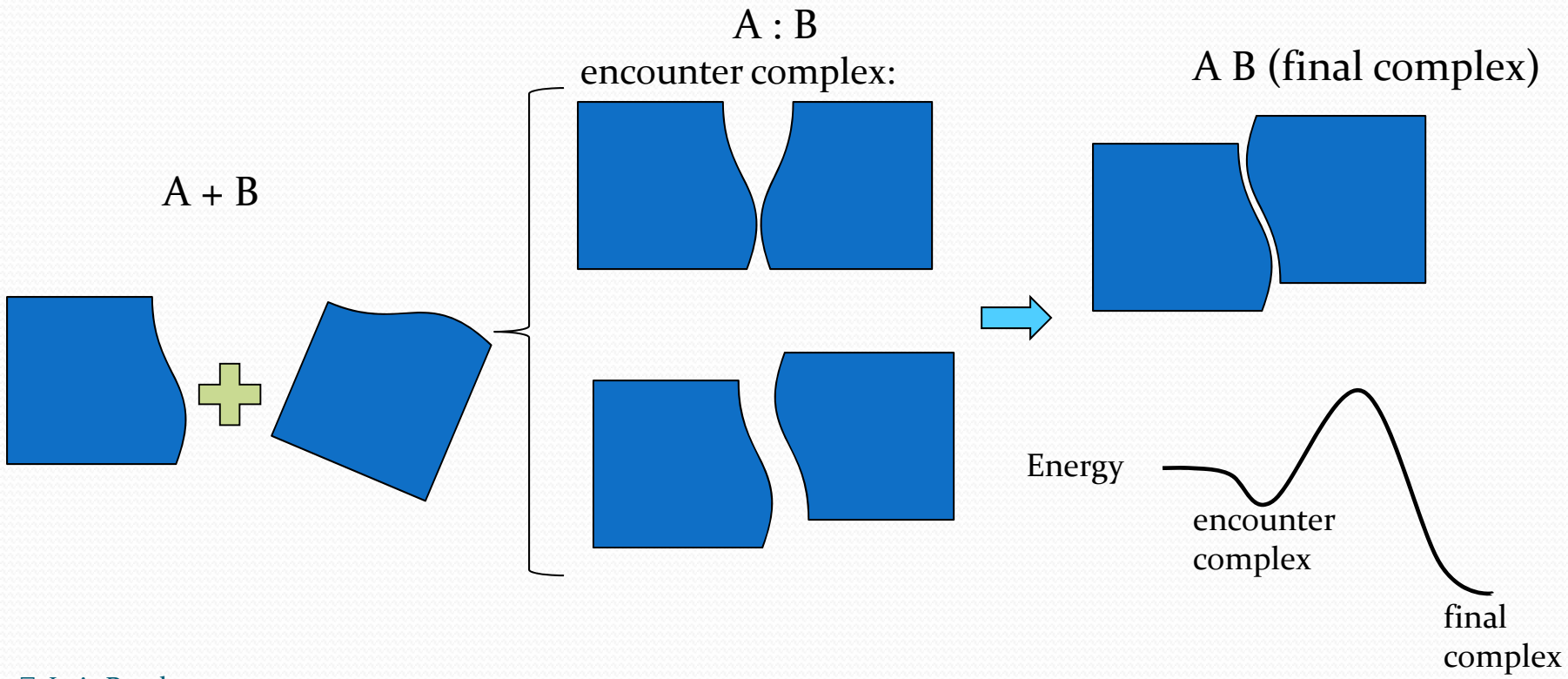


Properties Used in PPI Prediction

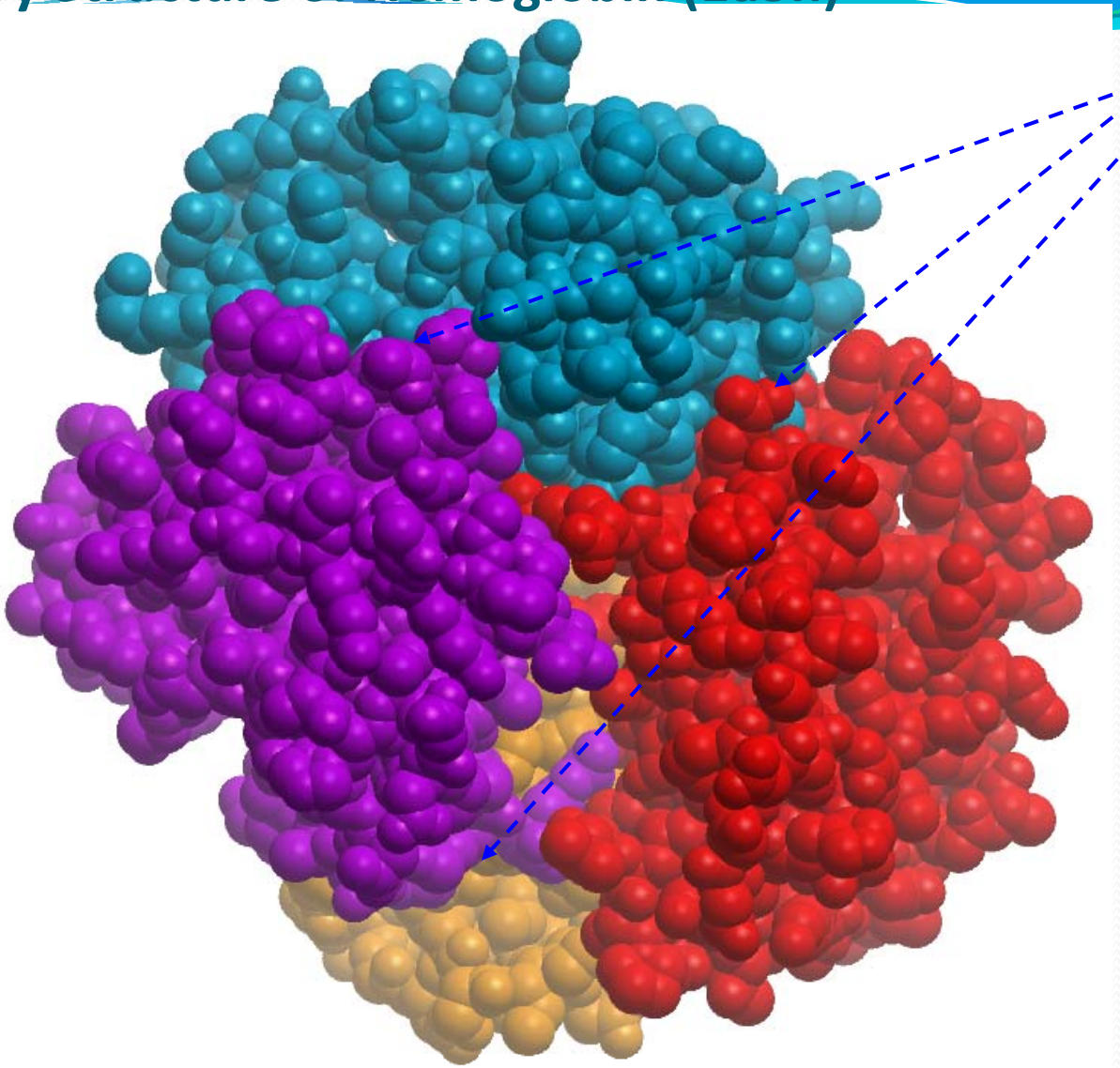
- Neighbor residues in spatial vicinity
- B-factor (flexibility of the residue)
- Solvent accessibility (solvent accessible surface area)
- Relative solvent accessibility
- Geometric: shape index, curvedness, planarity
- Predicted/approximate strucural features
- Evolutionary features: conservation scores, sequence profiles
- Physicochemical features
  - Hidrophobicity
  - Electrostatic potential
  - Desolvation energy
- External knowledge:
  - Interactions with other complexes
  - PPI interaction networks

Transient vs Obligate

- Association reaction of two proteins viewed as a random process.
- Collision at exact orientation lead to complex formation.
- Two proteins collide with each other at a rate given by diffusion to form a complex.
- Each collision forms an “encounter” complex (called transient).
- Precise collision will lead to final complex (called obligate)



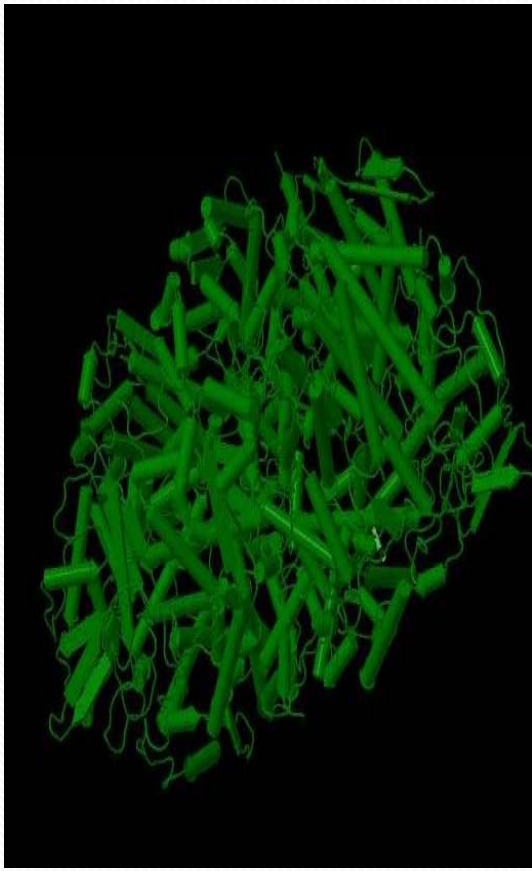
Quaternary Structure of Hemoglobin (1a3n)



- Interactions can last for *short* or *long* periods of time.

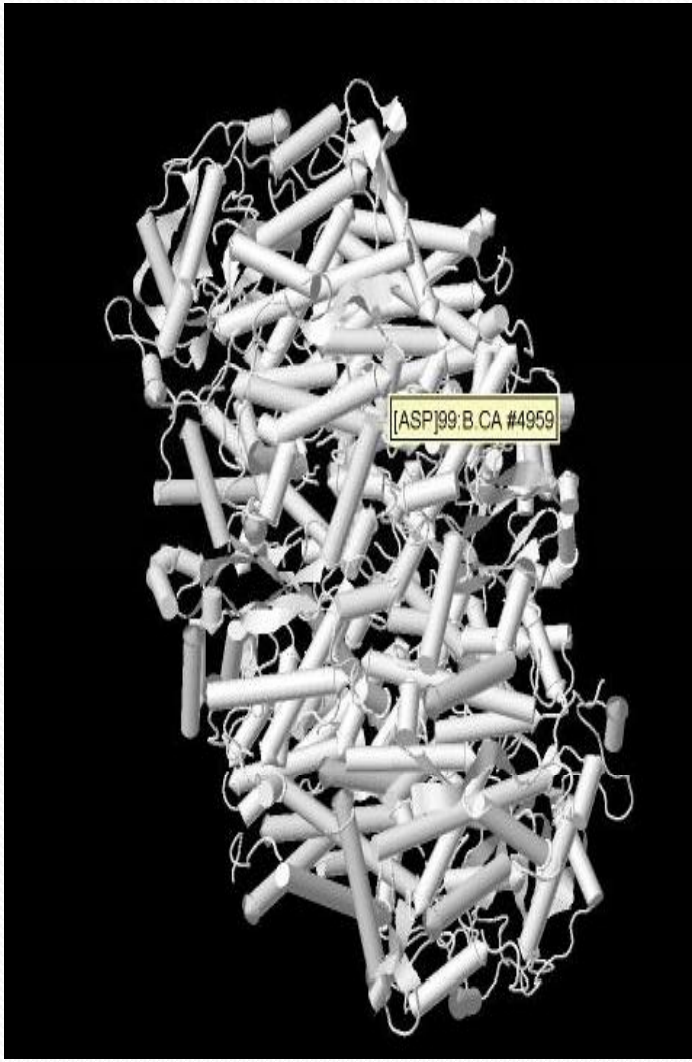
Protein-protein Interaction

- One of the current goals is to map the protein interaction networks into different organisms.
- In the complex web of interacting proteins to define a protein by its position needs protein-protein interaction information.
- Here, we work with individual complexes.
- We have proposed an approach to discriminate between two types of PPI interaction:
  - transient and obligate complexes



Protein-protein interaction previous work

- Interaction based on the lifetime of the complexes.
- Obligate: More stable
- Transient: Less stable and hence more difficult to discriminate.
- Zhu et al.: differentiate between Obligate and transient interactions based on solvent accessible surface area, conservation scores and shapes of the interface.
- Mintseris et al.: Based on atomic contact vectors



Binding Free Energy features: FastContact

- Each complex is listed in the form of one or more chains for ligand and receptor respectively.
- Feature Extraction was performed with FastContact.
- FastContact obtains a fast estimate of the binding free energy based on a **statistically determined solvation contact potential** and **Couloumb electrostatics** with a distance-dependent dielectric constant (4r).

FastContact Cont'd

For each complex: FastContact delivers:

- Total binding free energy
- Electrostatic energy
- Desolvation Free energy
- Top 20 max. and min. values of:
  - ✓ Residues contributing to the binding free energy
  - ✓ Ligand residues contributing to the desolvation free energy
  - ✓ Ligand residues contributing to the electrostatic energy
  - ✓ Receptor residues contributing to the desolvation free energy
  - ✓ Receptor residues contributing to the electrostatic energy
  - ✓ Receptor-ligand residue desolvation energies
  - ✓ Receptor-ligand residue electrostatic energies
- All these values(with the residue numbers) compose a total of 642 features.

Protein-protein interaction classification Contd.

- Classification: Given a complex **x**
  - A linear algebraic operation **y** = **A**x is applied.

    x is an *n*-dimensional vector.

    y is a *d*-dimensional vector where $d<<<n$

    A is a transformation matrix obtained from one of the LDR methods.
  - Vector **y** is then passed through:
    - Quadratic Bayesian classifier
    - Linear Bayesian classifier

Linear Dimensionality Reduction

- Applied three diffierent LDR methods:
  - Homoscedastic
    - Fisher's Linear Discriminant Analysis (FDA) (traditional LDA method)
  - Heteroscedastic
    - Loog-Duin Discriminant Analysis (HDA) (proposed by Loog and Duin in 2004)
    - Chernoff-based Discriminant Analysis (CDA) (proposed by Rueda and Herrera in 2008)

Experiments

- Two pre-classified, curated datasets of protein complexes were used.

- Dataset 1: Mintseris et. al. (MID)

  209 Transient complexes, 115 Obligate complexes.

- Dataset 2: Zhu et.al. (ZHD)

  62 Transient complexes, 75 Obligate complexes.

Subsets of Features

- To study the effects of the different types of energies and ligand/receptor,
- 13 different subsets of features for each dataset including: all 282 values (residue numbers were discarded)
  - binding free energies
  - ligand/receptor desolvation/electrostatic energies
  - ligand-receptor desolvation and electrostatic energies
  - desolvation and electrostatic energies.

Classification accuracy for MID dataset

Subset	n	k-NN	SVM	QB								
				FDA	SD	d^*	HDA	SD	d^*	CDA	SD	d^*
All Energetic	282	76.38	77.30	70.38	6.12	1	77.50	8.68	4	76.87	7.2	9
Binding Free Energy	40	69.94	72.09	71.86	7.09	1	75.20	10.77	6	73.33	17.56	5
Ligand Energy	80	72.70	74.54	66.58	11.18	1	76.42	7.55	8	76.44	8.74	6
Ligand Solvation	40	77.91	75.46	69.65	8.57	1	75.81	6.89	2	74.86	6.95	8
Ligand Electrostatic	40	69.33	70.86	72.09	9.16	1	72.14	6.46	7	71.84	11.31	4
Receptor Energies	80	74.54	74.23	67.17	9.85	1	76.42	6.46	6	76.74	6.36	11
Receptor Solvation	40	75.46	75.46	68.73	7.42	1	75.50	7.97	1	74.60	7.53	3
Receptor Electrostatic	40	72.09	70.55	68.47	8.81	1	69.71	4.84	3	70.31	6.53	12
Ligand-Receptor Energies	80	71.78	71.78	67.91	10.48	1	75.94	5.88	7	75.32	7.51	7
Ligand-Receptor Solv.	40	72.09	70.55	65.64	6.82	1	71.84	8.44	9	72.76	9.28	4
Ligand-Receptor Elect.	40	73.62	74.85	72.78	8.5	1	75.48	7.03	20	75.50	6.29	13
Solvation	120	78.53	76.07	65.72	6.53	1	76.70	9.18	14	76.41	9.52	11
Electrostatic	120	71.78	71.17	65.72	9.56	1	76.70	8.83	14	76.41	7.32	11

Classification accuracy for ZHD dataset

				QB								
Subset	n	k-NN	SVM	FDA	SD	d^*	HDA	SD	d^*	CDA	SD	d^*
All Energetic	282	67.15	65.69	58.62	6.12	1	65.08	8.68	1	69.12	7.2	15
Binding Free Energy	40	64.96	59.85	55.59	14.09	1	65.74	14.25	9	63.23	8.35	7
Ligand Energy	80	68.61	69.34	60.05	14.41	1	70.60	15.33	15	72.08	12.45	5
Ligand Solvation	40	70.80	70.80	62.35	13.41	1	70.64	14.45	18	69.26	16.82	6
Ligand Electrostatic	40	64.23	62.77	49.55	16.63	1	60.51	14.65	4	59.58	6.36	19
Receptor Solvation	40	76.64	64.96	66.05	11.31	1	74.03	17.54	11	73.97	10.08	13
Receptor Electrostatic	40	61.31	64.96	54.95	10.37	1	65.48	7.42	6	67.48	11.4	5
Ligand-Receptor Energies	80	67.15	67.88	67.22	11.09	1	69.16	13.51	17	70.97	15.56	5
Ligand-Receptor Solv.	40	70.8	70.07	70.71	8.18	1	72.08	14.29	10	72.18	8.37	18
Ligand-Receptor Elect.	40	61.31	55.47	60.27	12.25	1	66.72	11.93	16	67.54	10.38	17
Solvation	120	73.72	71.53	51.41	13.02	1	65.33	11.03	6	75.41	12.13	7
Electrostatic	120	69.34	72.99	53.61	15.23	1	63.53	8.71	14	64.10	9.58	1

Results for MID Dataset

- Best overall performance: LDR combined with QB classifier.
- Among them LDR criterion HDA achieves the best performance in 6 out of 13 cases.
- Classification of all LDR methods achieves the best performance in 10 out of 13 cases.
- k-NN achieved the best for desolvation energies of ligand-receptor.

Results on the ZHD Dataset

- Best performance: LDR combined with QB classifier.
- Among them LDR criterion CDA achieves the best performance in 8 out of 13 cases.
- Classification of all LDR methods achieves the best performance in 9 out of 13 cases.
- *k*-NN again gave the best in one of the datasets, desolvation receptor.

Analysis for Both Datasets

- A comparison with other subsets, desolvation energies achieves the best performance.

- suggests desolvation energies are the best properties to distinguish transient and obligate complexes for both datasets.

- Desolvation energy for ligand-receptor:
  - MID dataset: Best classification achieved by *k*-NN is 78.53%.
  - ZHD dataset: Best classification accuracy achieved by CDA is 75.41% while reducing from dimension 120 to 7.
  - In both cases, even better than using all features.

Conclusion

- Proposed a classification approach for transient and obligate protein-protein complexes.
- Used LDR that involve homoscedastic and heteroscedastic criteria coupled with a Quadratic Bayesian classifier.
- LDR schemes coupled with QB achieves best overall performance even better than K-NN and SVM with an RBF kernel.
- k-NN gives the best performance on individual subsets of features, involving desolvation energies.
- Comprehensive study shows that the best classification performance is achieved by using desolvation energies.
- The approach also considers (manual) feature selection.

Future work

- To use this approach in different protein-protein interaction classifications:
  - Intra and inter domains
  - Homo and hetero-oligomers.
- To use other features in this classification approach:
  - Solvent accessibility
  - Residual vicinity
  - Shape of the structure of the interface
  - Secondary structure
  - Planarity
  - Conservation scores
  - Other physiochemical features
  - Hydrophobicity
- To use automatic feature selection to infer individual properties, atoms and/or amino acids for characterizing PPIs.

Acknowledgments

- NSERC, the Natural Sciences and Engineering Research Council of Canada
- Canadian Foundation for Innovation
- Ontario Innovation Trust
- University of Windsor: Office of Research Services

Questions