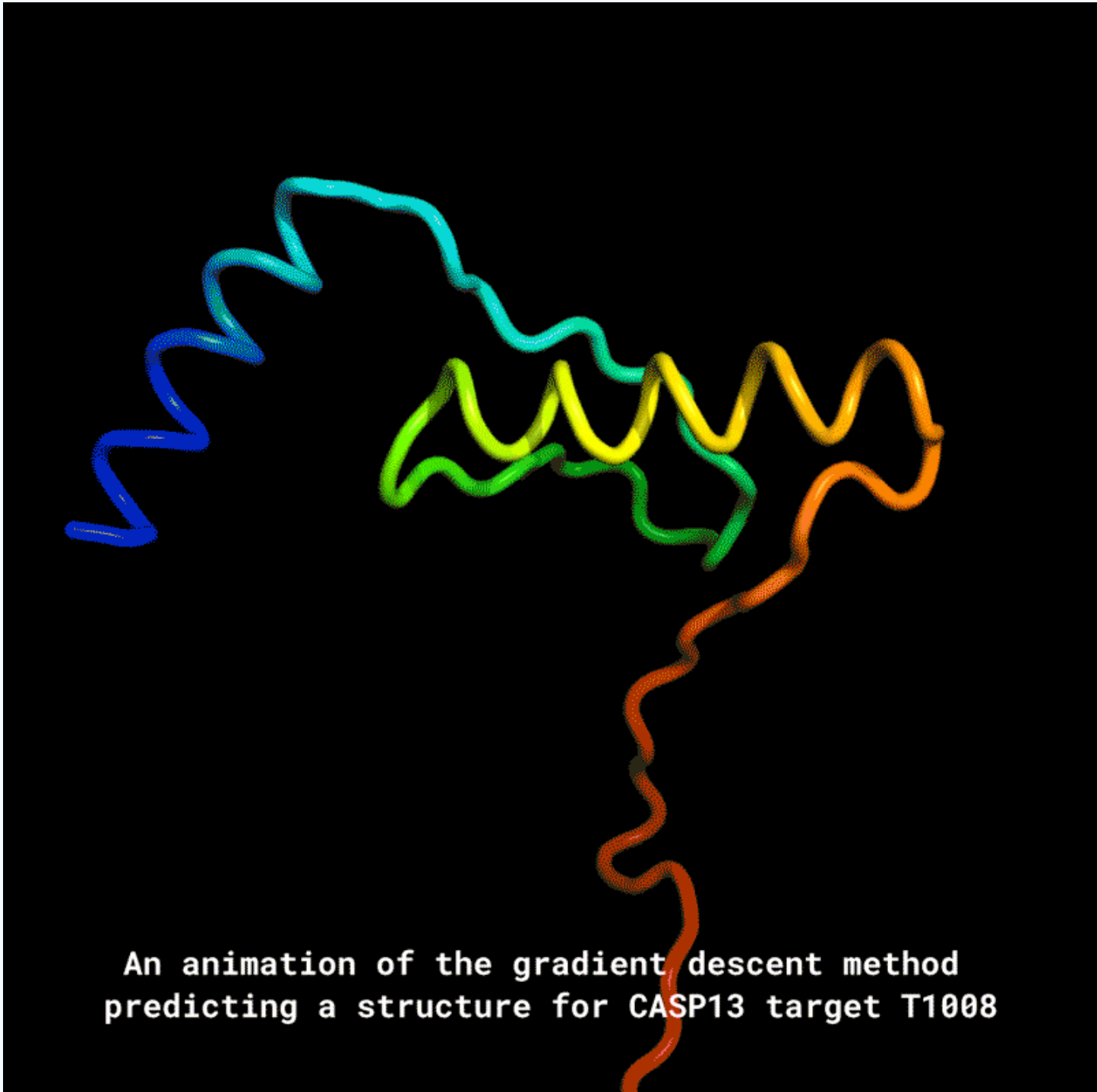


[AlphaFold, 2020]

Data Science

2020/06/17 Course 4

Applications of Information Science and Data Science in Biology

Data Science Center, Naoaki ONO, Ai Muto

Contents

1. Basic Statistics
2. Machine Learning: Classification, Clustering
3. Machine Learning II: PCA, regression, sequential data analysis, deep learning
4. **Applications of Information Science and Data Science in Biology**
5. Systems Biology
6. Descriptors in Material and Molecular Design
7. PCA・PLS (development of organic materials) and Pareto solutions (Catalyst design)

Contents

1. 基礎統計
2. 機械学習I:分類、クラスタリング
3. 機械学習II:PCA、回帰、系列データ学習、深層学習
4. **バイオサイエンスにおけるビッグデータ解析とデータサイエンス**
5. システムバイオロジー
6. マテリアル・分子設計における記述子
7. PCAとPLS(有機材料の開発)、パレート最適解(触媒設計)

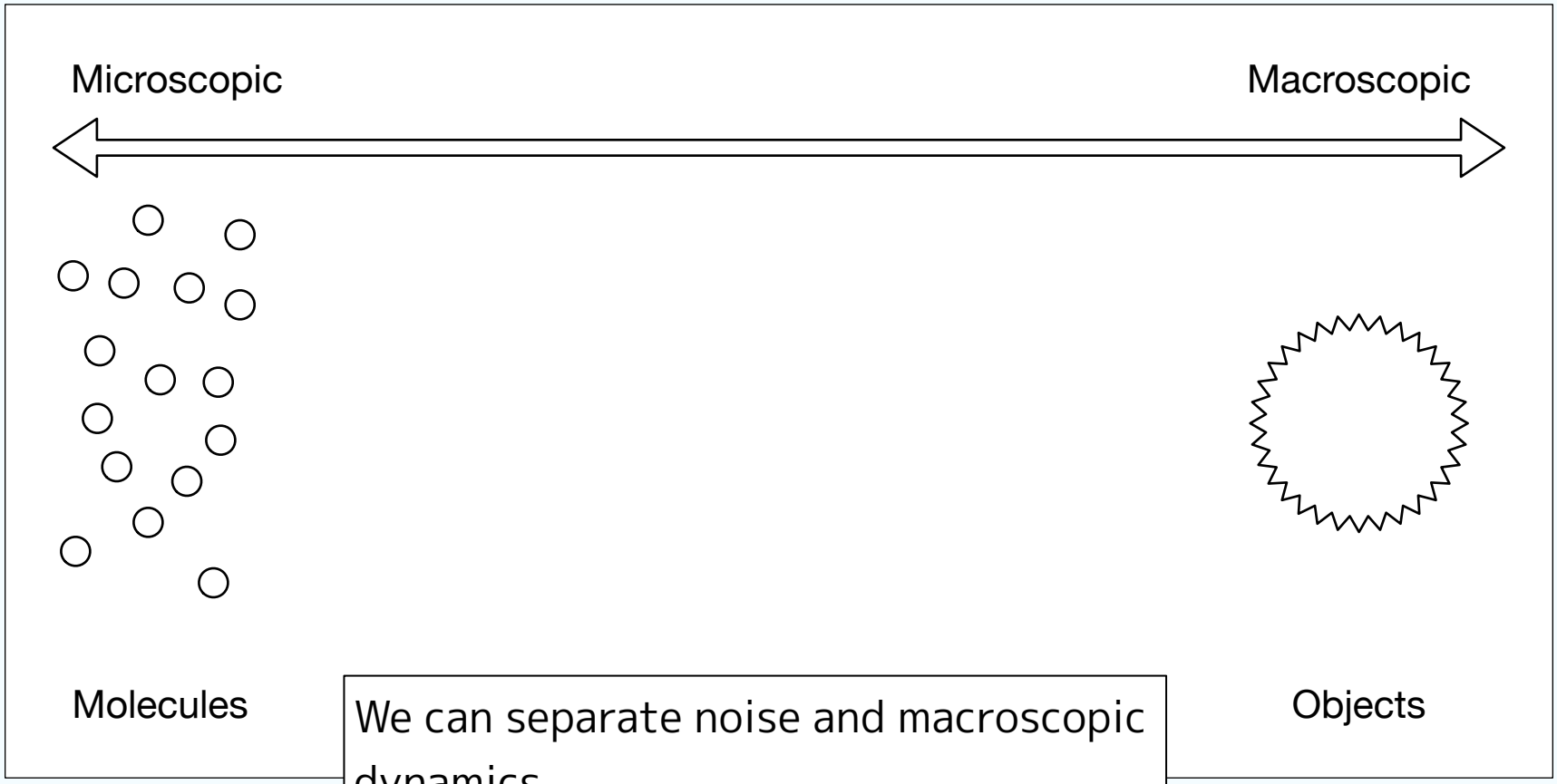
Preface

- Biology and Informatics
  - Why informatics required in biology?
  - Why modeling biological system is so difficult?
  - How can we improve likelihood of biological model?

Complex Systems

Complex Systems

- **Classical Physics**



We can separate noise and macroscopic dynamics.

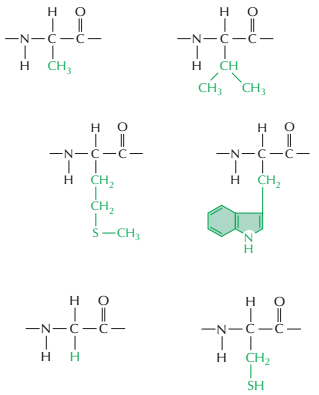
$$\sum f_{\mathrm{mol}}(\mathrm{v}(t)) \simeq \epsilon_{\mathrm{noise}}(t) + f_{\mathrm{object}}(\mathbf{x}(t))$$

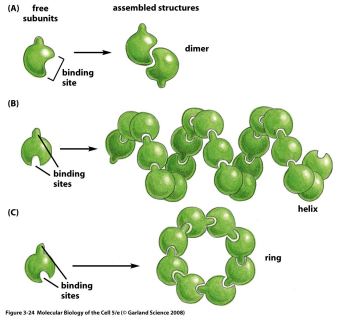
Complex Systems

- **Complex Systems**

Microscopic

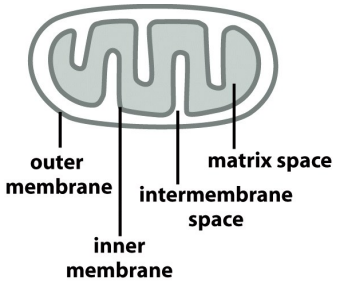
Macroscopic



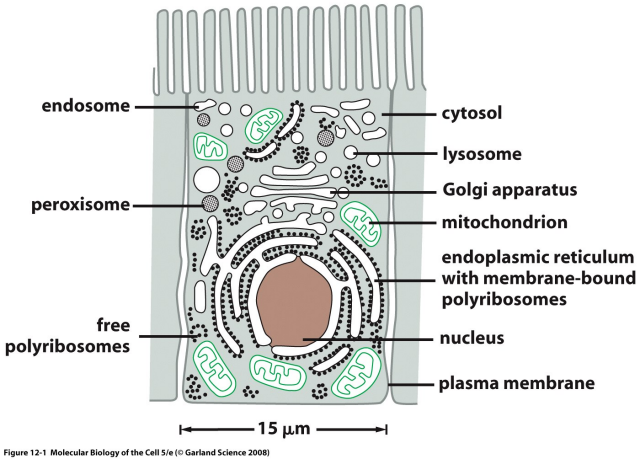


Molecules

Polymers



Organelle



Cells

We can **not** separate microscopic and macroscopic dynamics.

$$\simeq \epsilon_{\text{noise}}(t) + \sum f_{\text{polymers}}(\mathbf{v}(t)) \times f_{\text{organelle}}(\mathbf{w}(t)) \times f_{\text{cells}}(\mathbf{x}(t))$$

Complex Systems

- **Classical biology**
  - Classification and qualitative description
  - Focus on the behavior where number of interactions are very small.
- **Complex biology**
  - Where number of interaction are large, we a huge number of parameters to describe the dynamics.
  - If we have very **huge data samples**, and **appropriate physical model**, we may able to describe and predict their behavior.

Enrichment Analysis

Q. Do you think how many mutations do you have in your genome?

Q. Do you think how many mutations do you have in your genome?

A. 4 ~ 5 million sites

> There ain't no "normal person".

Enrichment Analysis

- **Polymorphisms**
  - 99.8 % of our genome (~30Gbp) are identical.
  - Variation shared by some population (e.g. >1%) are called "polymorphisms".
  - Most of polymorphisms are Single Nucleotide Polymorphisms (SNPs).

A random mutation which did not spread.

```
ATGCAGATCGACTAGCGTA
ATGCAGATCGACTAGAGTA
ATGCAGATCGACTAGCGTA
ATGCAGATGGACTAGCGTA
ATGCAGATGGACTAGCGTA
```

A polymorphism shared by a certain amount of population

Enrichment Analysis

Though we can read gene sequences, it does not implies that we can see how it works.

- **Genome-Wide Association Study (GWAS)**
  - Statistical analysis to find association between genetic variants and phenotypic traits such as human diseases.

> Towards prediction model of human traits from his/her genome data...?

Enrichment Analysis

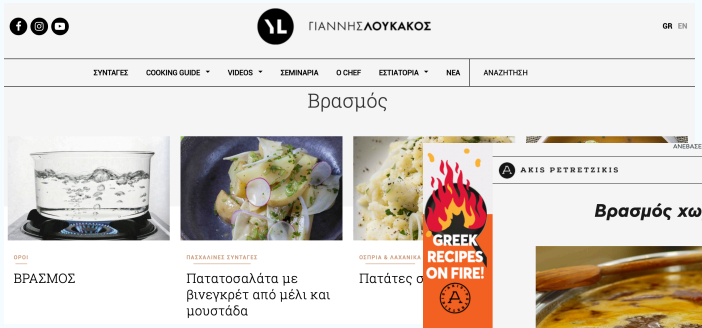
- **Example**

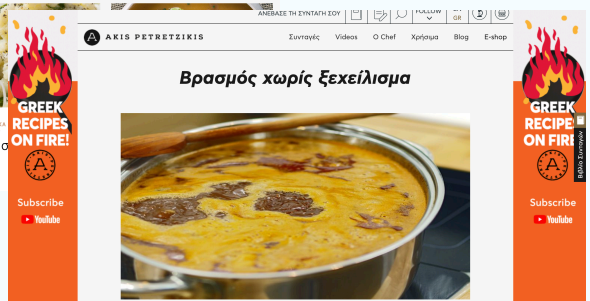
How you can find the meanings of the greek word

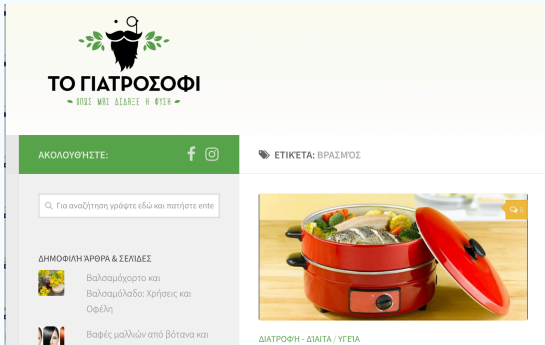
βρασμός

without dictionaries?

> If you have a bunch of texts that include this words and you know that most of that texts are related with "cooking", you can associate this word with cooking.







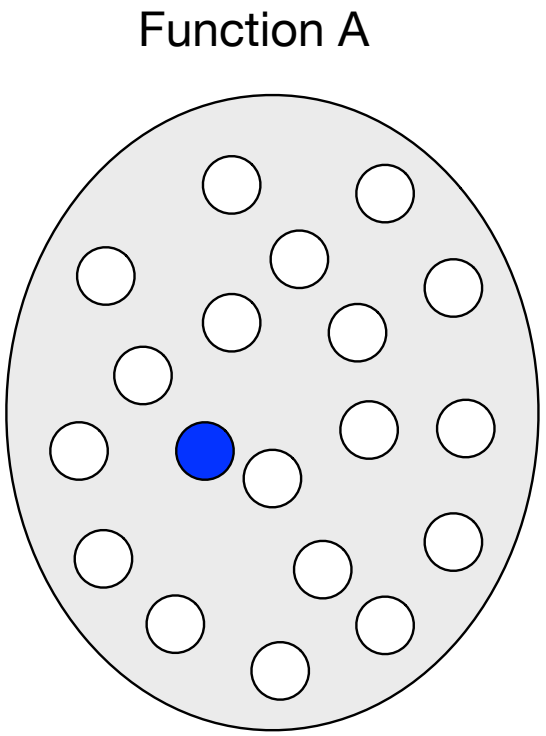
Enrichment Analysis

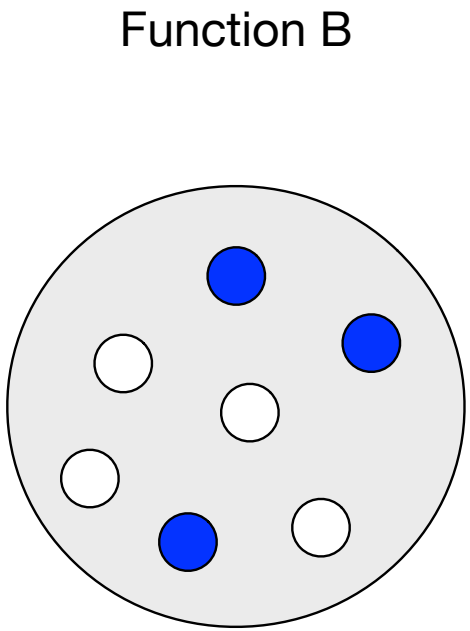
- **Gene Set Enrichment Analysis**
  - You don't know which genes are related with **DiseaseX.**
  - You have a list of genes where SNPs are significantly often observed in the genomes of patients of DiseaseX.
  - You have a list of gene set "**Functional Categories**" in which most of the human genes have been categorized.

  - > If SNPs are more often appeared in **FC-α**, compared with other FC, it suggest that DiseaseX is caused by some problems with the **Function-α.**

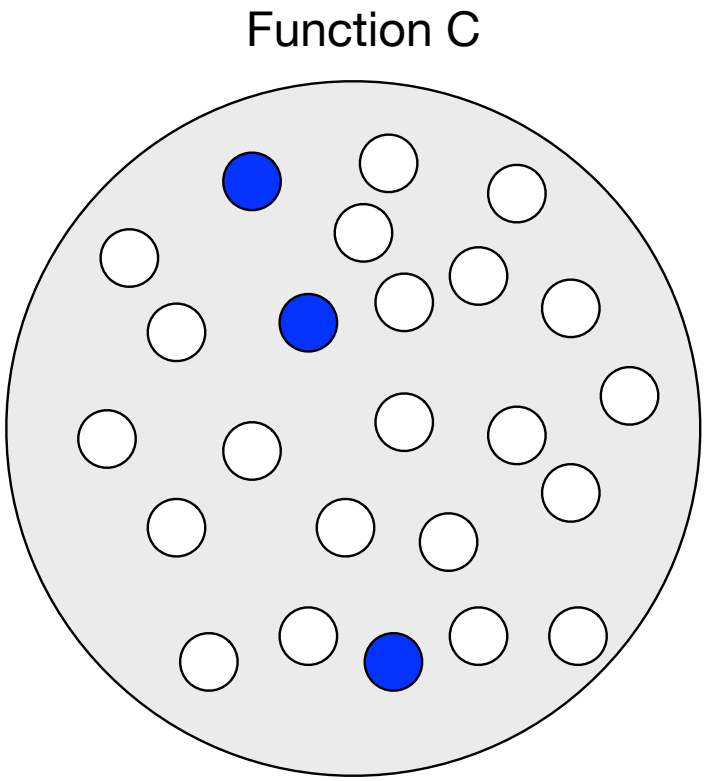
Enrichment Analysis

- **Hypergeometric test**





	A	B	C	Tot
SNP	1	3	3	7
WT	18	4	21	43
Tot	19	7	24	50



Three SNPs out of seven are significantly often or not

Enrichment Analysis

- **Hypergeometric test**
  - When the ratio of SNPs was K/N in total.
  - Randomly pick n genes and count the number of SNP.
  - Compute the probability that p_{Hyp} that k SNPs are picked out of n.
  - If $\sum_{i \geq k} p_{\mathrm{hyp}}(i, n)$ is smaller than a threshold τ, we claim that "SNPs are significantly enriched in this category".

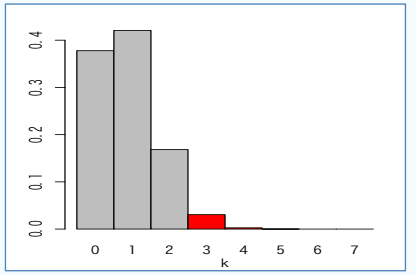
$$p_{\mathrm{hyp}}(k, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Enrichment Analysis

- **Hypergeometric test**
  - e.g., Function B: $n = 7, k = 3$

$$\sum_{i=3}^{7} p_{\mathrm{hyp}}(i, 7, N = 50, K = 7)$$
$$\simeq 0.04776071 < 0.05$$

	A	B	C	Tot
SNP	1	3	3	7
WT	18	4	21	43
Tot	19	7	24	50



"SNPs were significantly enriched ($p < 0.05$) in FC-B".

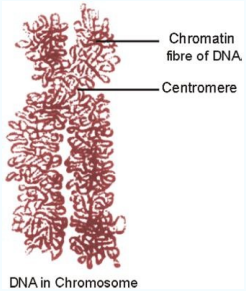
Thus we further investigated the relationship between the function B and DiseaseX ...

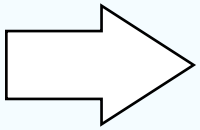
Note: Generally, the number of considering categories will be very large, so that we need to avoid over estimation due to "multiple comparison".

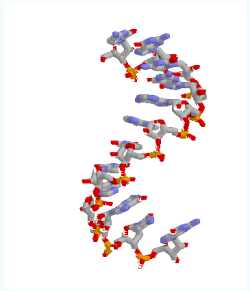
Multiomics Analysis

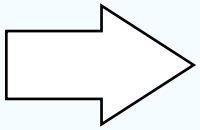
Information theory and Biology

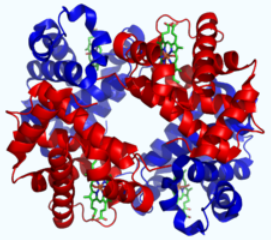
- **Static vs Dynamic**
  - Evolution of life has separated "**gene:information**" and "**enzyme:machine**".
  - It's consistent with architecture of current computers.



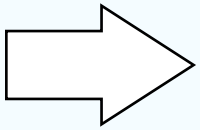


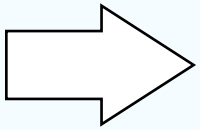


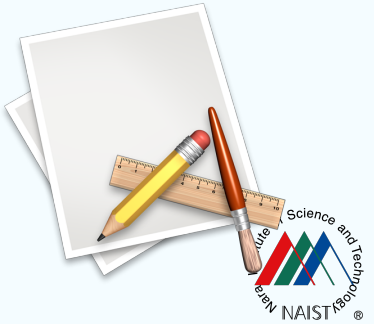




DNA/Static memory　　mRNA/dynamic memory　　protein/op codes







Multiomics Analysis

- **Genome**
  - Complete gene sequences
- **Transcriptome**
  - Comprehensive profiles of gene activity
- **Proteome**
  - Comprehensive profiles of synthesized proteins
- **Metabolome**
  - Comprehensive profiles of synthesized compounds
- **Phenome**
  - Variations of cell shapes, behaviors, and other traits.

Integrate some of these layers as a statistical model to gather more information and improve its likelihood.

Multiomics Analysis

- **Genome**
  - Homology search: find the best alignment allowing insertion / deletion in the query strings.

mismatch

insertion/deletion

> Define scores for **matches**, **mismatches** and **gaps** and find an alignment that minimize total penalty using **Dynamical Programming**.

- **Dynamical Programming**

QUERY:

str1: TGCTCGTA

str2: TTCATA

SCORE:

match: +5

mismatch: −2

indel: −6

Fill the matrix by adding the scores. It represents all possible alignments. Starting from right-bottom cell, traceback the "best" score up-leftwards.

[Nature Biotech. 2004]

Dynamic programming matrix:

		j → (sequence **y**)								
		0	1	2	3	4	5	6	7	8 = N
			T	**G**	**C**	**T**	**C**	**G**	**T**	**A**
i	0	0	−6	−12	−18	−24	−30	−36	−42	−48
	1 **T**	−6	5	−1	−7	−13	−19	−25	−31	−37
	2 **T**	−12	−1	3	−3	−2	−8	−14	−20	−26
	3 **C**	−18	−7	−3	8	2	3	−3	−9	−15
	4 **A**	−24	−13	−9	2	6	0	1	−5	−4
	5 **T**	−30	−19	−15	−4	7	4	−2	6	0
M = 6	**A**	−36	−25	−21	−10	1	5	2	0	11

(sequence **x**)

Optimum alignment scores 11:

T	**−**	**−**	**T**	**C**	**A**	**T**	**A**
T	**G**	**C**	**T**	**C**	**G**	**T**	**A**
+5	−6	−6	+5	+5	−2	+5	+5

Protein Folding

- A **sequence of amino acid** in a protein is determined by the codons in its gene sequence.
- 3D **folding of a protein** depends on the interactions between amino acids.
- The function of a protein determined by their folded structure.

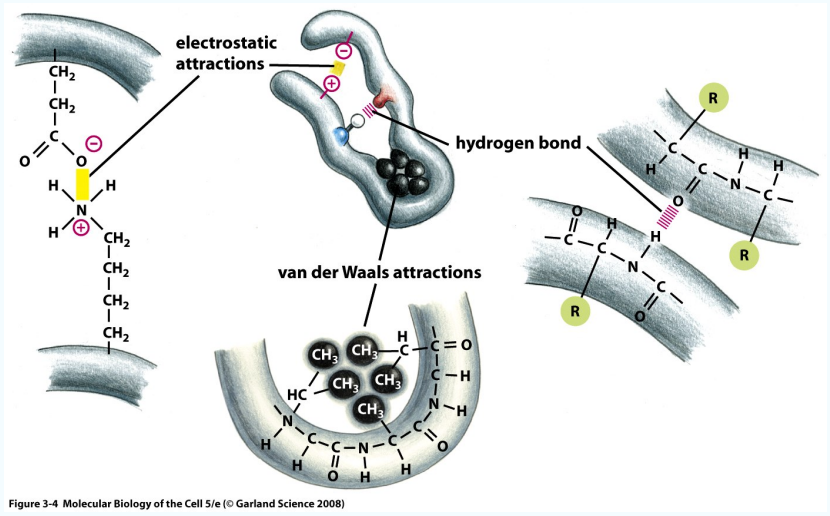


Figure 3-4 Molecular Biology of the Cell 5/e (© Garland Science 2008)

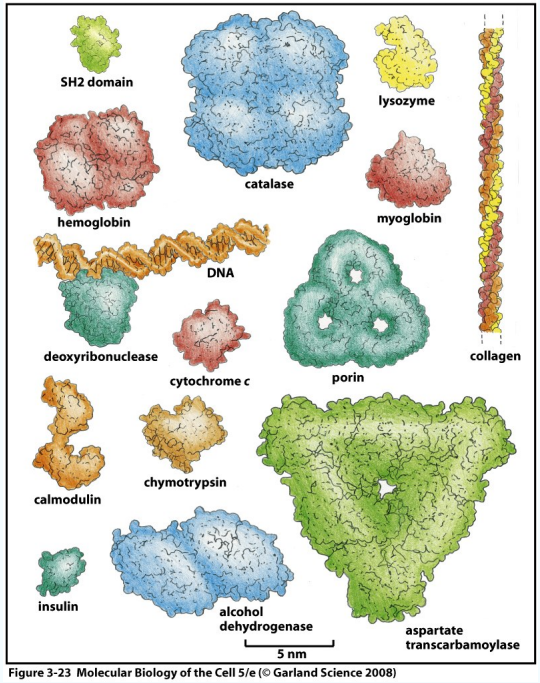


Figure 3-23 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Protein Folding

- **Prediction of protein folding**
  - We can know gene sequences, consequently, amino acid sequences.
  - Direct prediction of protein folding based on AA seq will require **molecular dynamics simulation of thousands atoms** using supercomputers.

> Can't we predict protein structures from AA seq more efficiently?

Protein Folding

- **AlphaFold**
  - Apply Deep Learning to optimize the parameters of physical model to estimate interaction between AA.
  - Compute probabilistic distribution of distance between two monomers using combined model of the DL model and physical features.
  - Maximize likelihood of folding structures using numerical optimization.

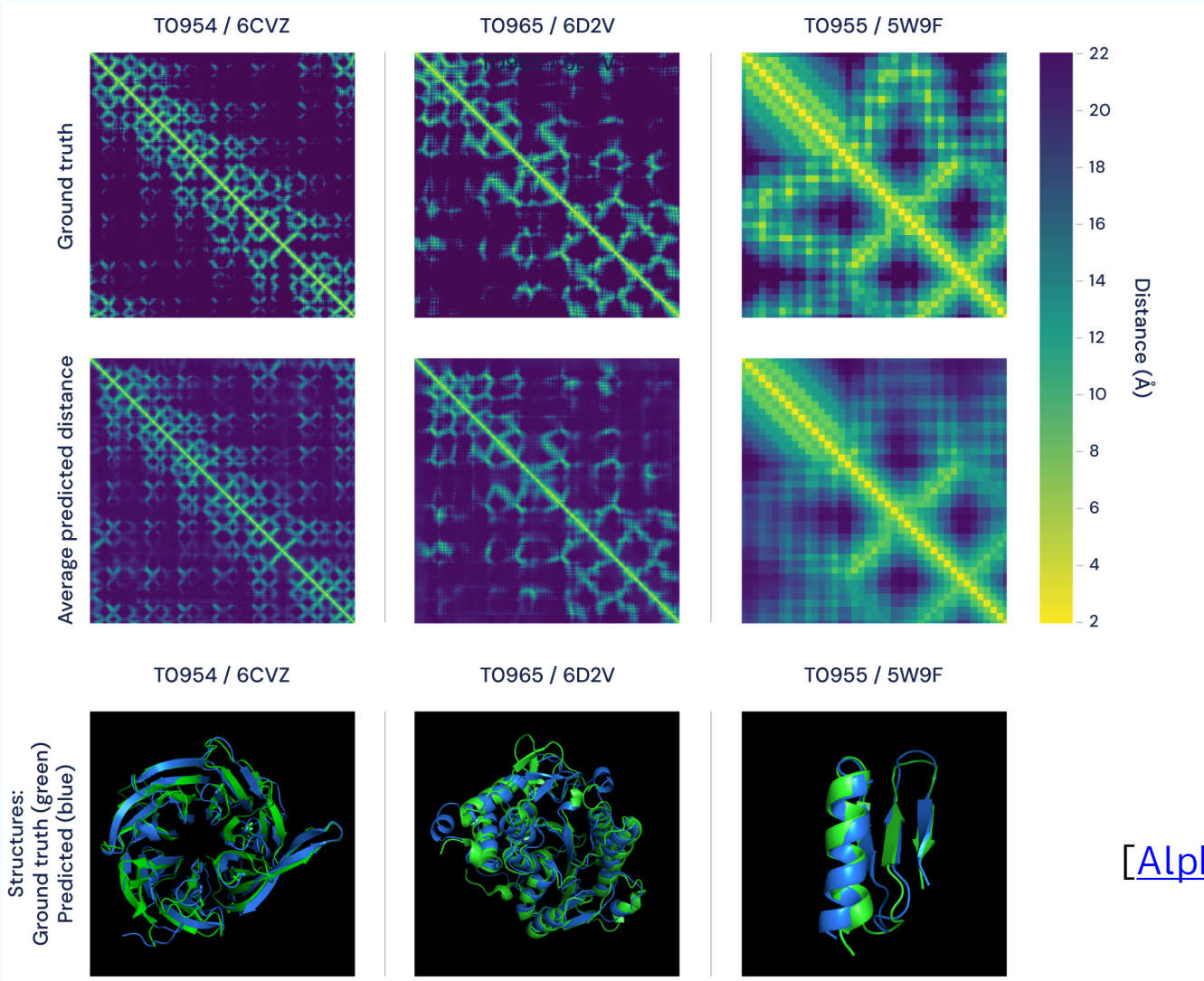
Protein Folding

- **AlphaFold**

Observed distance matrix

Predicted distance matrix

Observed & predicted structures



[AlphaFold, 2020]

Biophysical Modeling

DNA Hybridization

- **DNA microarray (photolithographic synthesis)**

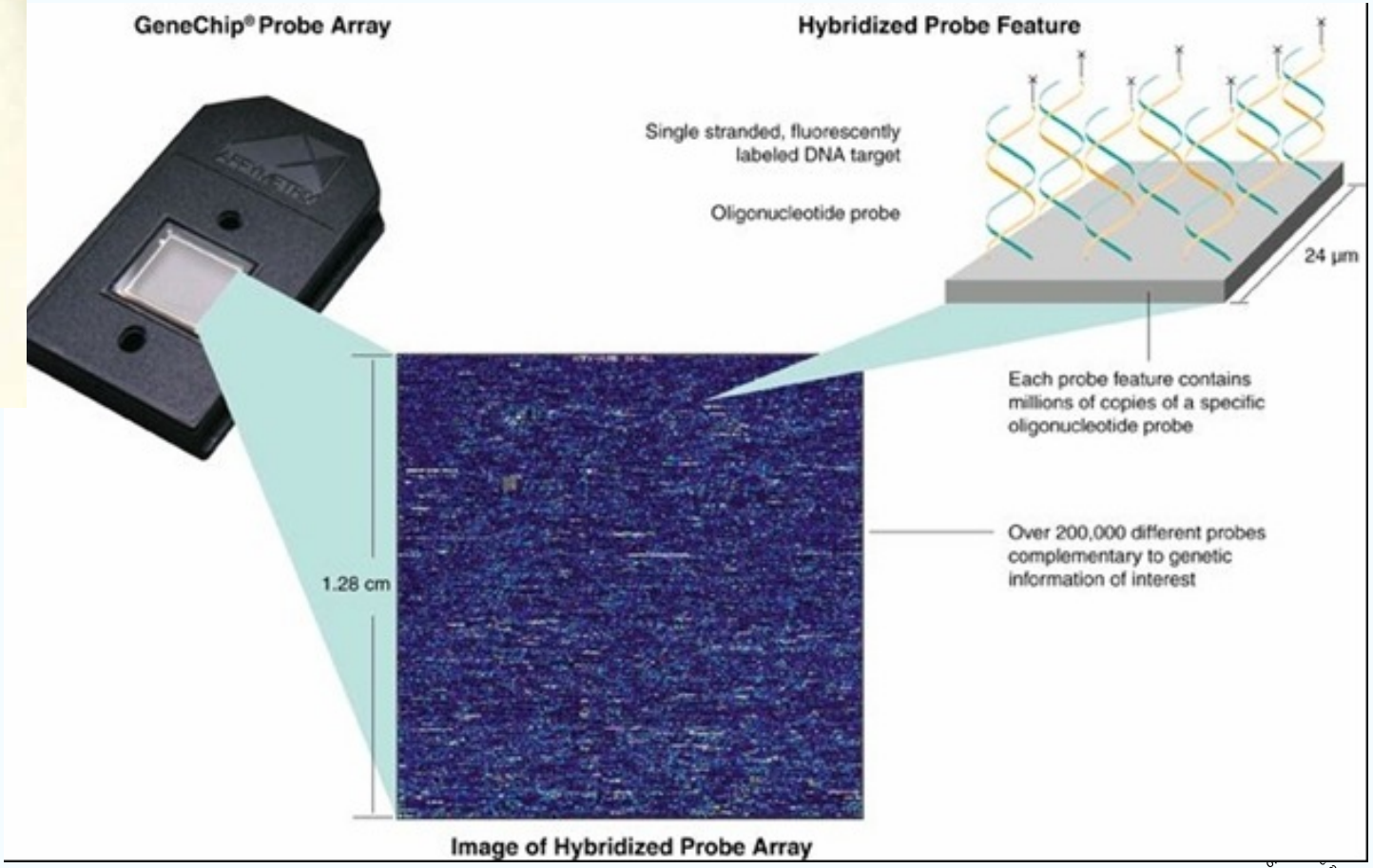
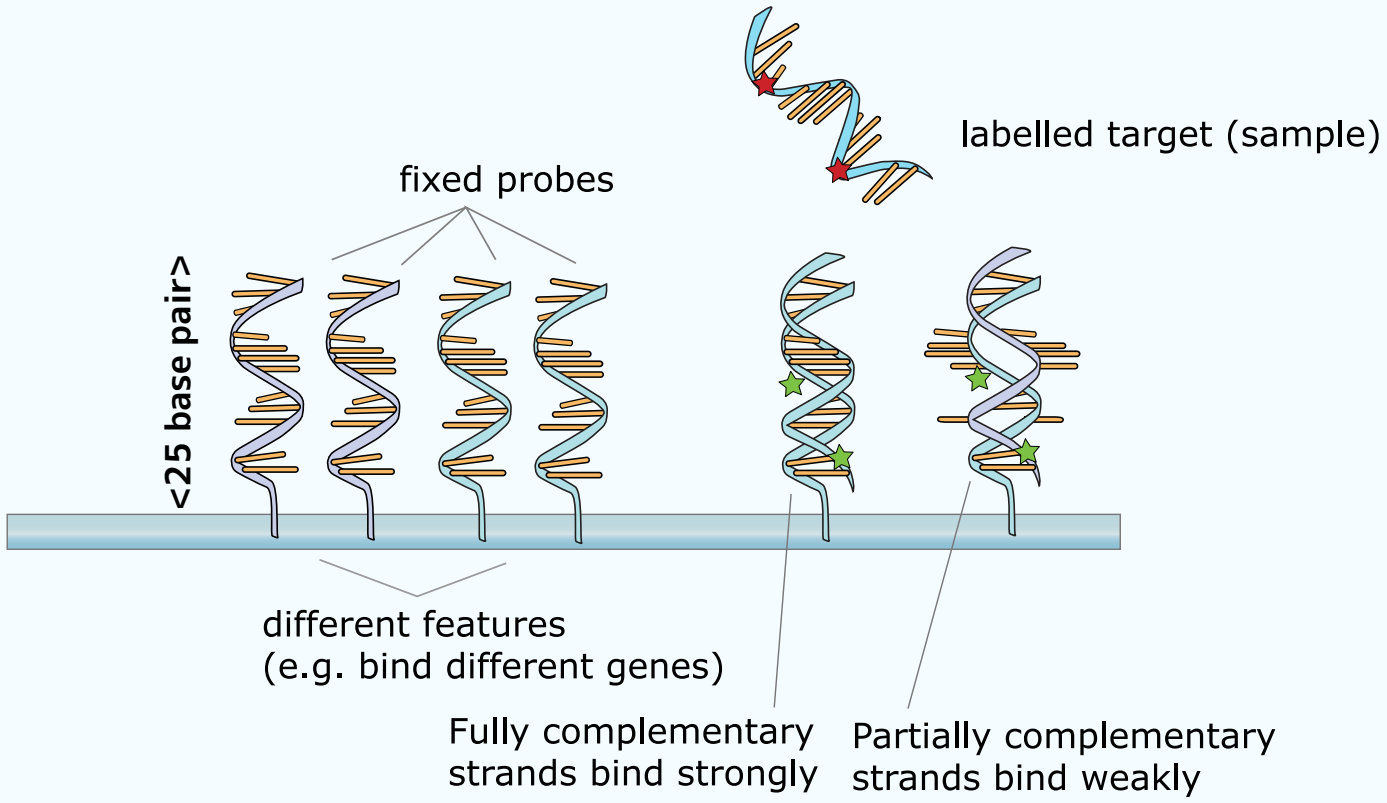


Figure. A design of an oligonucleotide microarray.

DNA Hybridization

- **DNA microarray**



Schematic illustration of DNA microarray. Designed probes are fixed on specific location of an array. When labelled targets were hybridized and detected, they are identified by their position of the signal.

DNA Hybridization

- **Quantification of expression values**
  - Model of hybridization on chip

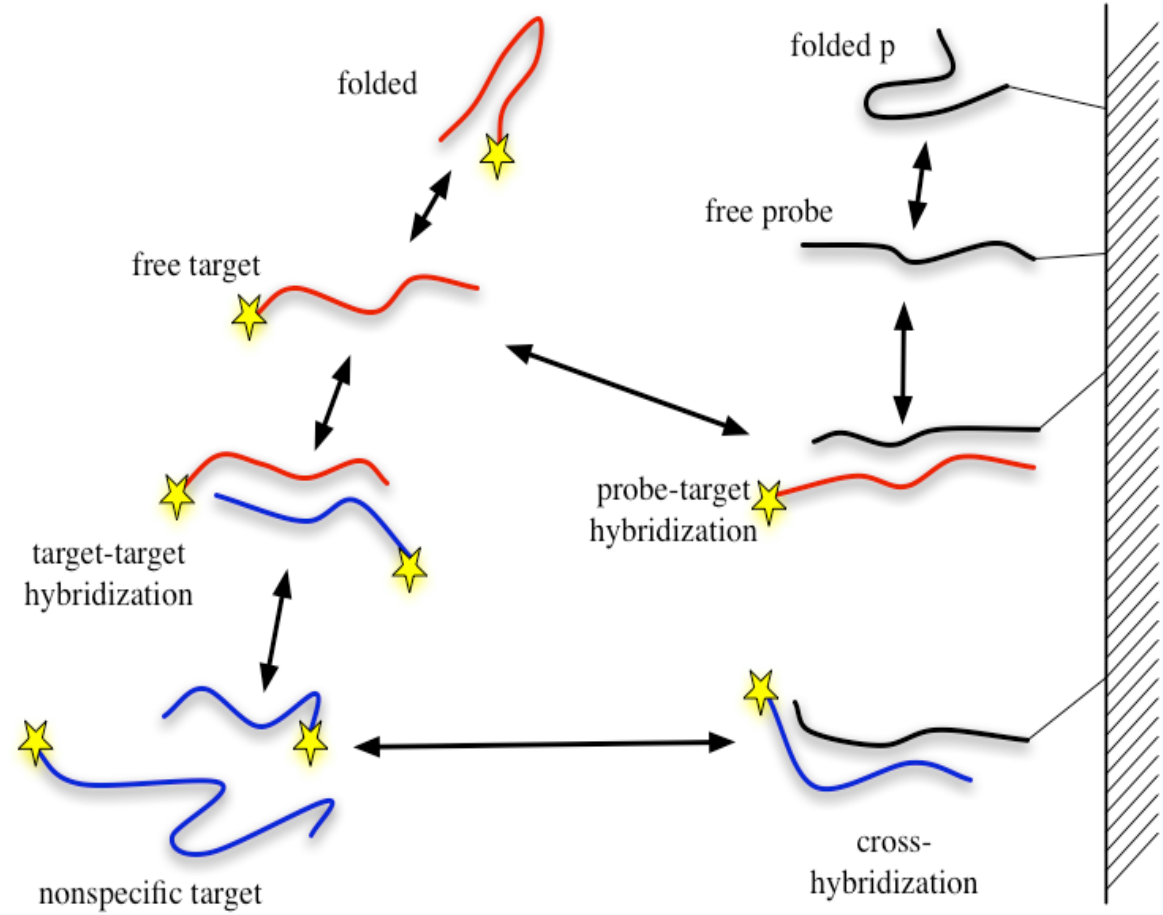


Figure. Model of hybridization. Equilibrium between labeled target DNA fragments in sample and probes synthesized on the microarray surface.

DNA Hybridization

- **Quantification of expression values**
  - Model of hybridization on chip

Intensity of a probe (I) is predicted as a function of probe sequence ($\mathbf{b}$), where target concentration is given by x. Effect of non-specific hybridization K^{ns} and self-folding K^{fold} are also taken into account. C, N, A and w are given model parameters.

$$\Delta G^{\mathrm{sp}}(\mathbf{b}) = \sum_{k=2}^{20} \{ w(k) \epsilon(b_{k-1}, b_k, b_{k+1}) \}, \quad (1)$$

$$I(\mathbf{b}) = \log_{10} \left(C(X^{\mathrm{sp}}(\mathbf{b}) + X^{\mathrm{ns}}(\mathbf{b})) + I^{\mathrm{bg}} \right) \quad (2)$$

$$X^{\mathrm{sp}}(\mathbf{b}) = \frac{1}{2} \Big\{ 1/K^{\mathrm{eff}}(\mathbf{b}) + A + x \quad (3)$$

$$- \sqrt{\left(1/K^{\mathrm{eff}}(\mathbf{b}) + A + x\right)^2 - 4Ax} \Big\}$$

$$K^{\mathrm{eff}}(\mathbf{b}) = \frac{K^{\mathrm{sp}}(\mathbf{b})}{1 + K^{\mathrm{fold}}(\mathbf{b}) + K^{\mathrm{ns}}(\mathbf{b}) N} \quad (4)$$

$$X^{\mathrm{ns}}(\mathbf{b}) = \frac{\left(A - X^{\mathrm{sp}}(\mathbf{b})\right) K^{\mathrm{ns}}(\mathbf{b}) N}{1 + K^{\mathrm{fold}}(\mathbf{b}) + K^{\mathrm{ns}}(\mathbf{b}) N}, \quad (5)$$

DNA Hybridization

- **Validation of the prediction**

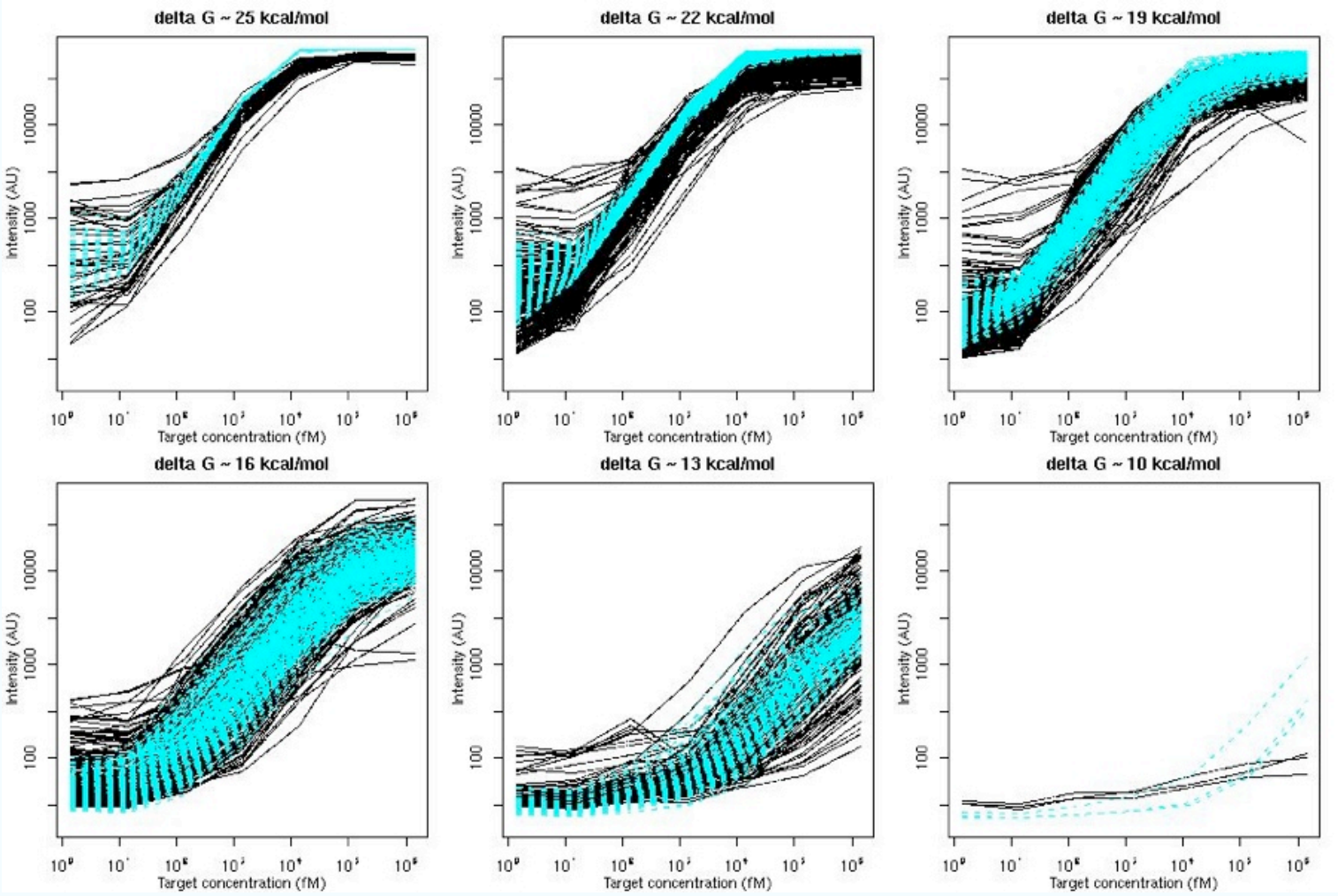


Figure. Intensity as function of target concentration. Black and blue lines represents observed and predicted intensity.

Afterwords

- **Biology and Informatics**
  - **Why informatics required in biology?**

    Shift from qualitative to quantitative description, and predict their behavior.
  - **Why modeling biological system is so difficult?**

    In biological system, it is difficult to separate microscopic and macroscopic dynamics.
  - **How can we improve likelihood of biological model?**

    Introduce more parameters based on biophysics and optimize them a large amount of experimental data.

Report

	A	B	C	Tot
SNP	1	3	3	7
WT	18	4	21	43
Tot	19	7	24	50

- **Task1**

  Test if FC-C is significant or not by the hyperbolic test,

  i.e., compute the probability $p_{\mathrm{hyp}}(k \geq 3, n = 24, K = 7, N = 50)$

- **Task 2**

  Find the best alignment of the following query and show the total matching score.

  QUERY:

  str1: CGATAGTTA, str2: AGTAGCTTC

  SCORE:

  match: +5, mismatch: −2, indel: −6

Filename: 4_{StudentID}_{LastName}_{FirstName}.pdf

Deadline: 2020/07/03