KU LEUVEN

KERNEL METHODS FOR GENOMIC DATA FUSION

POOYA ZAKERY

Centre for Brain and Disease Research, Flanders Institute for Biotechnology (VIB), Leuven and Department of Neurosciences and Leuven Brain Institute, KU Leuven, Leuven, Belgium

VIB
science meets life

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

A challenge in computational biology

- Gigantic bottleneck: finding a protein's structure and function given its sequence
  - There is an increasing gap between protein sequence information and protein structural and functional information

Computational Biology

Machine Learning Approach

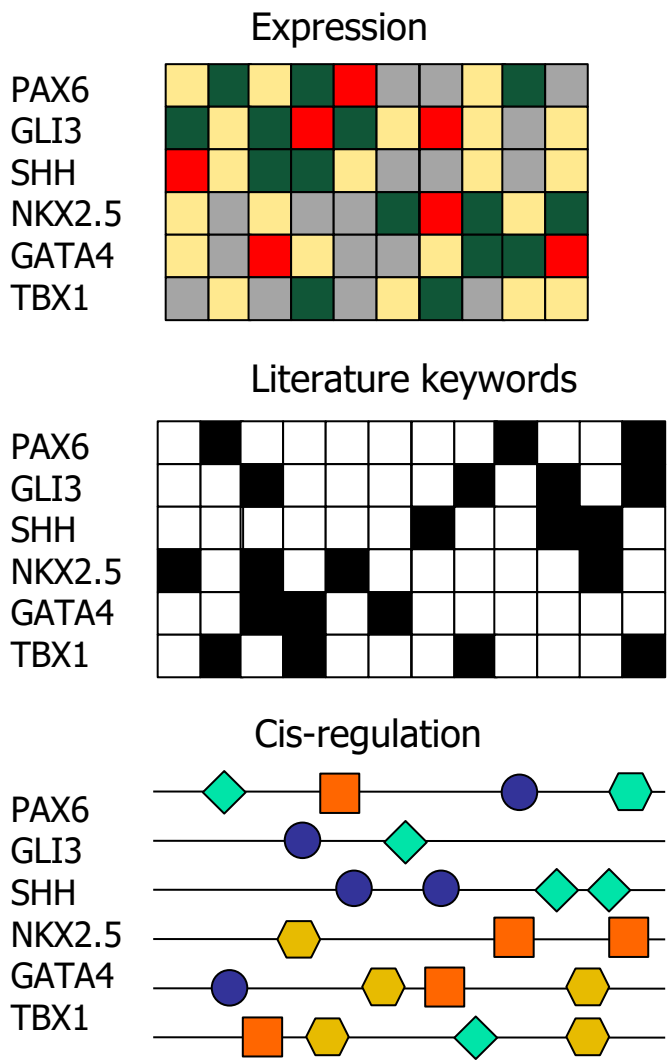
A challenge in computational biology

- Protein annotation database: Incomplete, inconsistent, biased, inaccurate, and even incorrect annotations are difficult to work with and can easily propagate
- Mosaic development of biology
- **Better understanding and modeling of biological systems**

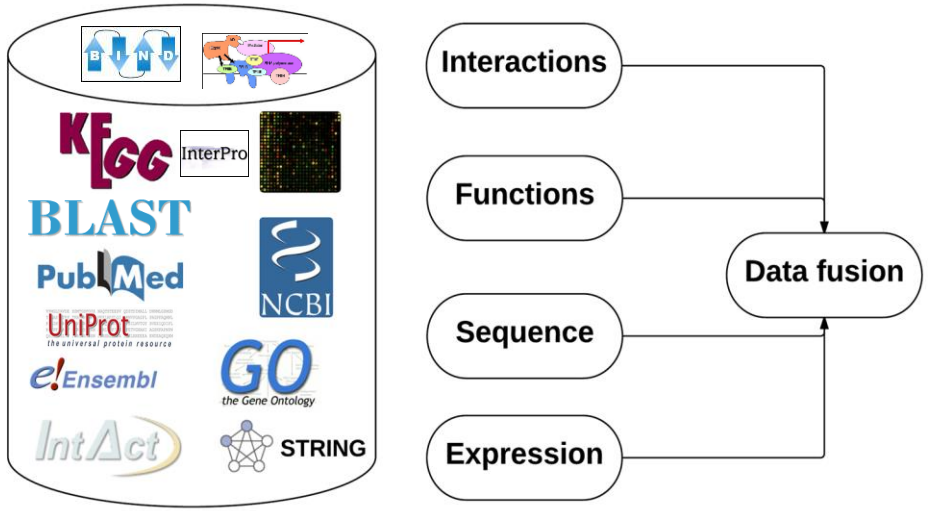
Genomic Data Fusion

Machine Learning Approach

Other challenges



- Biological data sources are heterogeneous by their very nature
- Large-scale data
- **Kernel Method** is particularly an interesting approach when there are heterogeneous sources (representations) of data for the task at hand (we consider classification and prioritization)



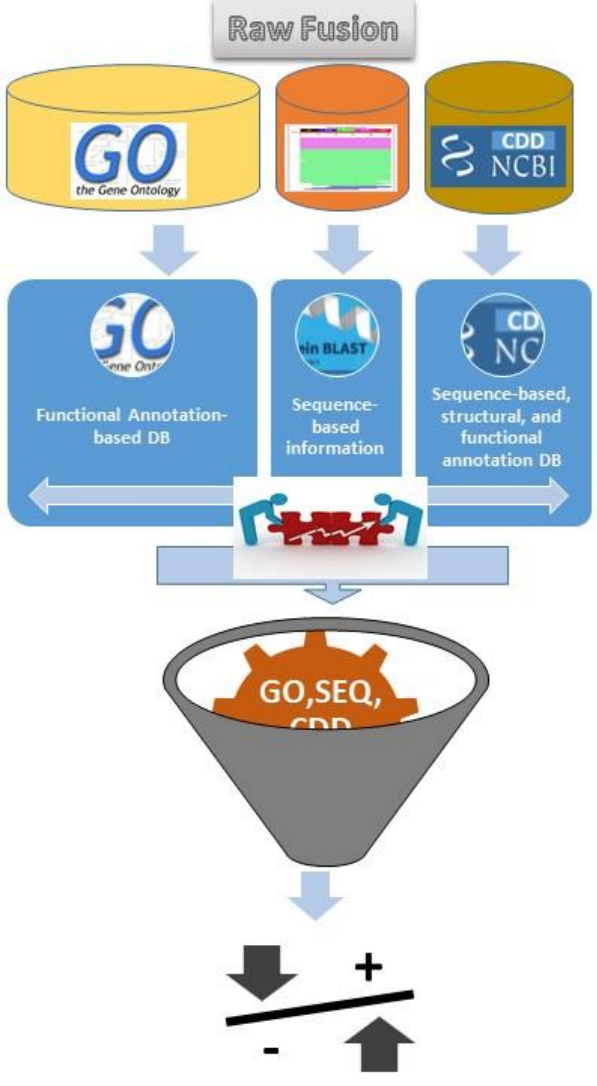
Scalable Machine Learning Approaches

KU LEUVEN

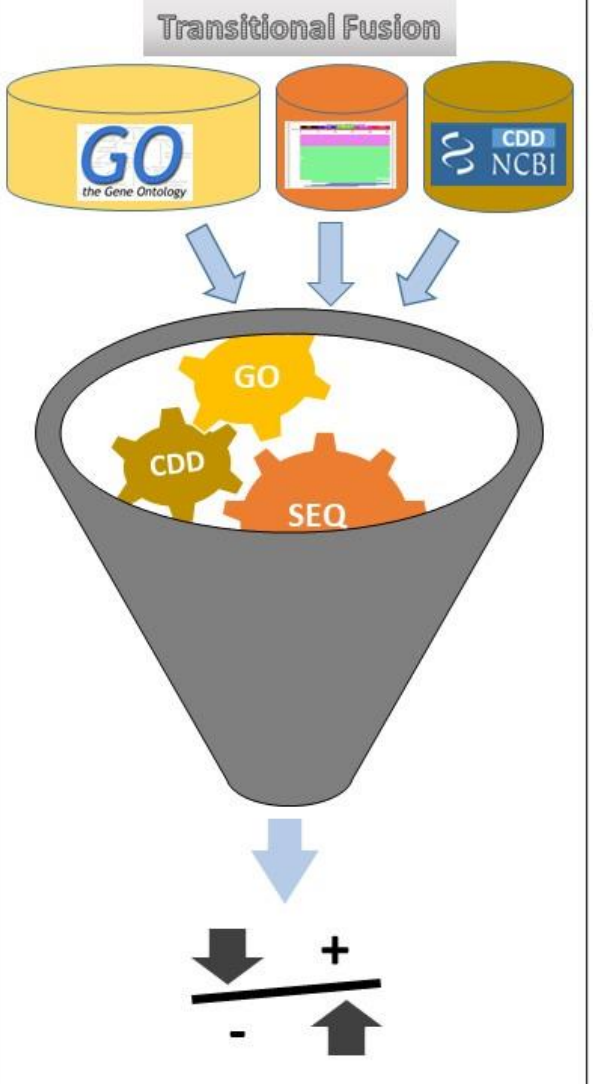
OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

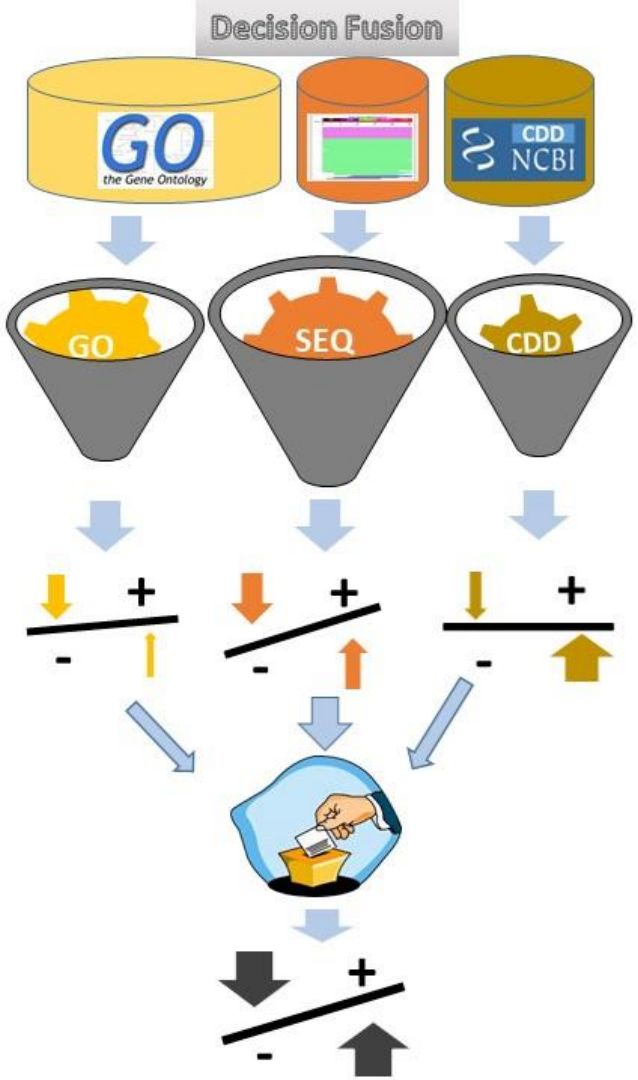
DATA FUSION SCHEMES



Bayesian row data fusion (GeneHound)
Gene prioritization,
Text mining



Geometric Kernel Fusion
Gene prioritization
Fold recognition
Protein localization



Late kernel integration
CAFA2, sub-mitochondrial localization

KU LEUVEN

DATA FUSION SCHEMES

Prediction of protein submitochondria locations based on data fusion of various features of sequences. J Theor Biol. 2011

Protein fold recognition using geometric kernel data fusion. Bioinformatics. 2014

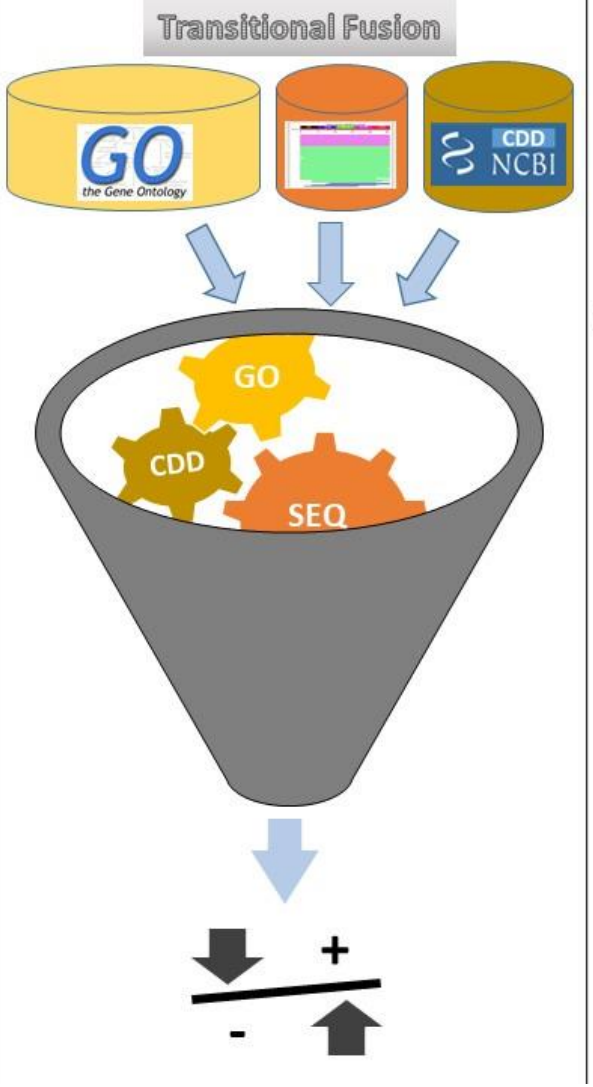
Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics. 2018

An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 2016 Sep

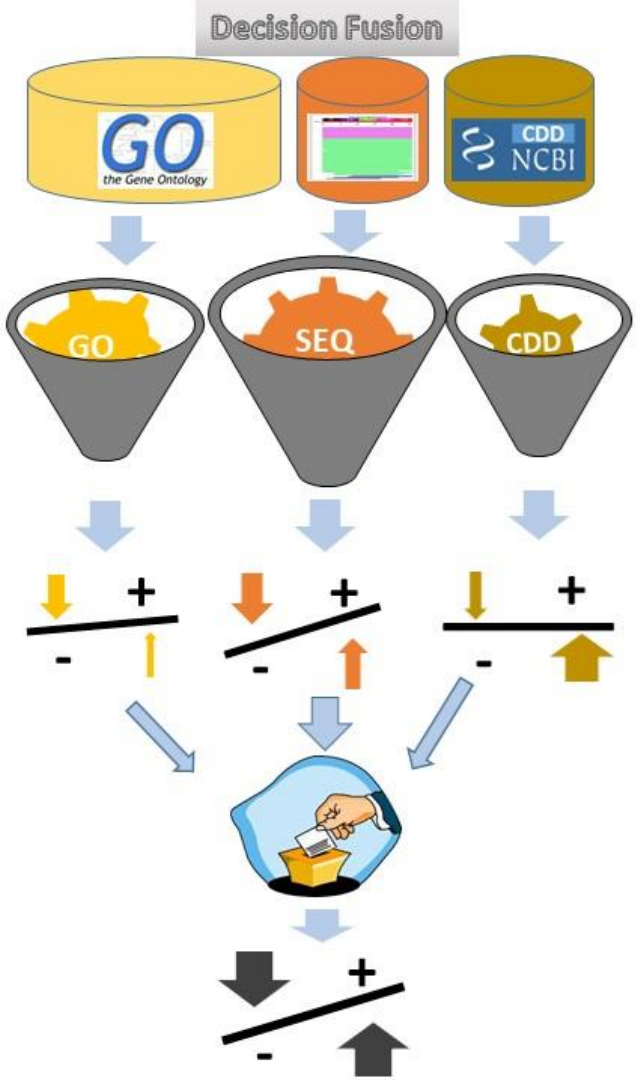
Application of geometric kernel data fusion in protein fold recognition and protein sub-nuclear localization. MLSB. 2014

Gene prioritization through geometric-inspired kernel data fusion IEEE/BIBM. 2015.

Bayesian row data fusion (GeneHound)
Gene prioritization,
Text mining



Geometric Kernel Fusion
Gene prioritization
Fold recognition
Protein localization

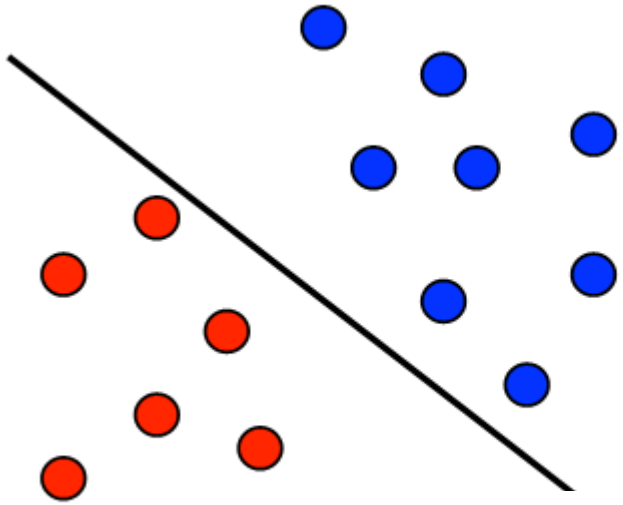


Late kernel integration
CAFA2, sub-mitochondrial localization

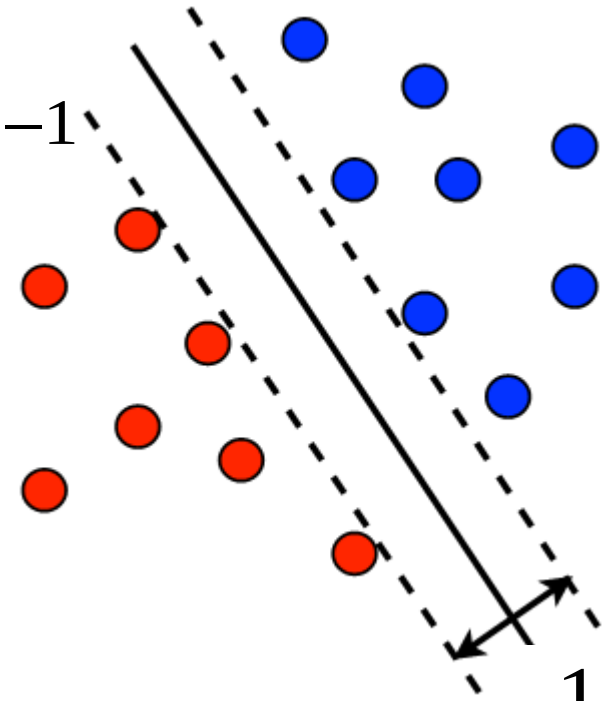
KU LEUVEN

Recap - SVM

$$f(x) = w.\phi(x) + b$$



$$w.\phi(x) + b = 0$$

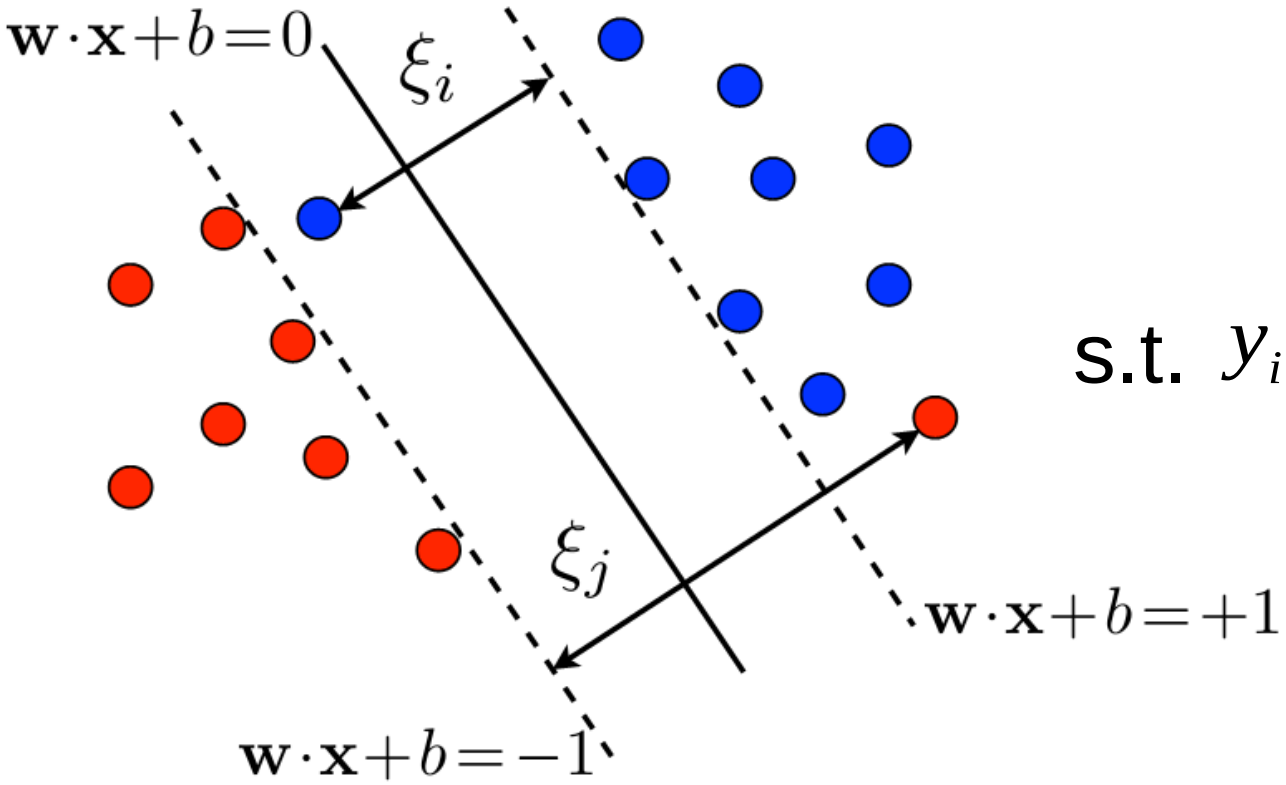


$$\frac{1}{\|w\|}$$

Maximize Margin

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$\textbf{s.t.} \quad y_i(w.\phi(x_i) + b) \geq 1 \quad \forall i$$



$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

s.t. $y_i(w.\phi(x_i)+b) \geq 1-\xi_i$

$\xi_i \geq 0 \ \forall i$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \max\left(0, 1-y_i(w.\phi(x_i)+b)\right)$$

Regularizer

Loss Function

$l(f(x_i), y_i)$

SVM: Optimization Problem: Primal vs. Dual

Primal

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

s.t. $$y_i(w.\phi(x_i)+b) \geq 1-\xi_i \quad \xi_i \geq 0 \quad \forall i$$

$$\Downarrow$$

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \left(\phi(x_i).\phi(x_j)\right)$$

s.t. $$0 \leq \alpha_i \leq C \quad \forall i$$

$$\sum_i \alpha_i y_i = 0$$

Dual

SVM-Dual

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \left(\phi(x_i).\phi(x_j) \right)$$

s.t. $0 \le \alpha_i \le C \quad \forall i$

$$\sum_i \alpha_i y_i = 0$$

$$K(x_i, x_j)$$

Classifier:

$$f(x) = w.\phi(x) + b = \sum_i \alpha_i y_i \left(\phi(x_i).\phi(x) \right) + b,$$

$$b = y_i - \sum_j \alpha_j y_j \left(\phi(x_i).\phi(x) \right)$$

Kernel Methods

Ideas:

Define $K: X \times X \to \Re$, called kernel, such that:

$$K(x, y) = \phi(x).\phi(y)$$

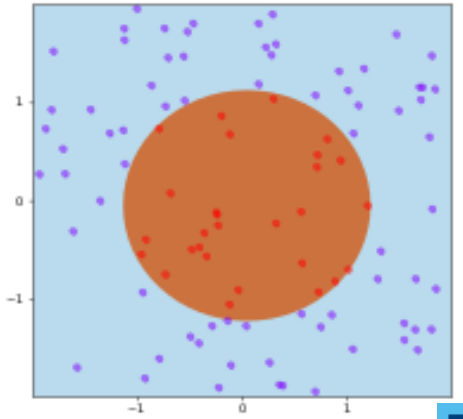
kernel matrix K often interpreted as a similarity measure

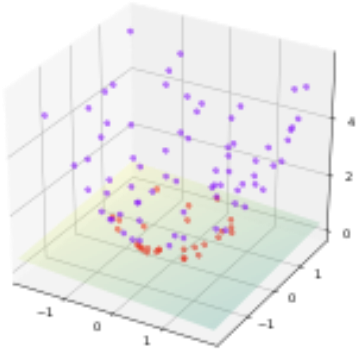
$$K(x, y) = (x_1 y_1 + x_2 y_2 + c)^2$$

$$= \begin{bmatrix} x_1^2 \\ x_1^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2c} x_1 \\ \sqrt{2c} x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} x_1^2 \\ x_1^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2c} x_1 \\ \sqrt{2c} x_2 \\ c \end{bmatrix}$$

Advantages:

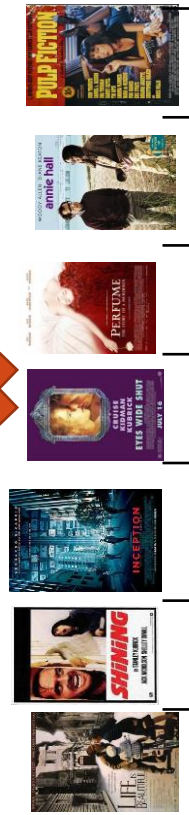
- Efficiency
- Flexibility





KERNEL METHODS

Pooya	5	6	9	7	3	9	8
Maria	7	8	10	5	**9**	10	9
Ghazal	4	6	9	7	5	6	8
Noah	9	8	8	8	8	9	5
Mohsen	5	6	8	2	8	7	9
Albert	9	6	8	10	6	8	8



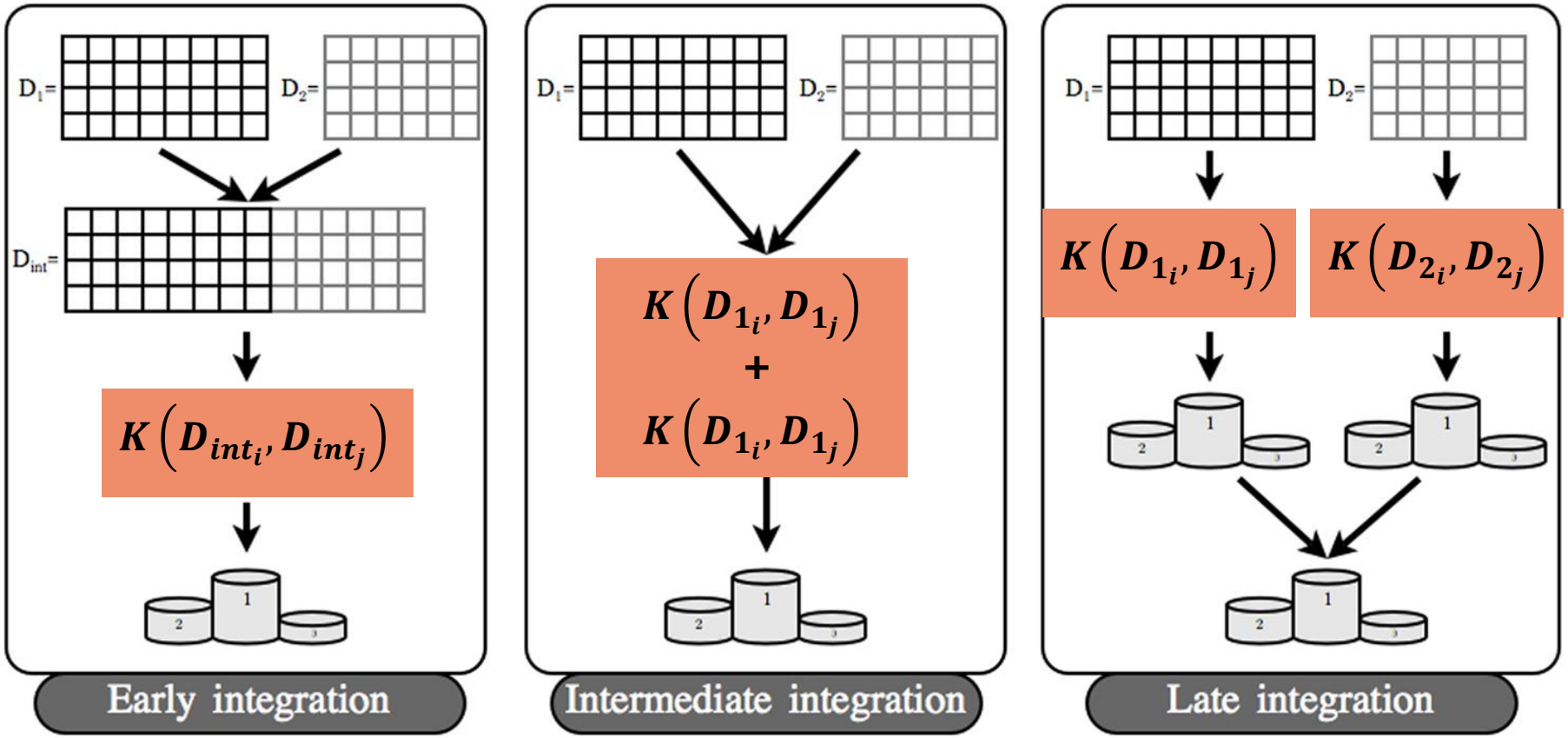
Pooya	Maria	Ghazal	Noah	Mohsen	Albert
5	7	4	9	5	9
6	8	6	8	6	6
9	10	9	8	8	8
7	5	7	8	2	10
3	9	5	8	8	6
9	10	6	9	7	8
8	9	8	5	9	8

	Pooya	Maria	Ghazal	NOah	Mohsen	Albert
Pooya	1	.97	.99	.96	.95	.98
Maria	.97	1	.98	.98	.99	.97
Ghazal	.99	.98	1	.96	.97	.98
Esan	.96	.98	.96	1	.95	.98
Mohsen	.95	.99	.97	.95	1	.94
Amin	.98	.97	.98	.98	.94	1

Kernel matrices are the nonlinear extension of covariance/correlation matrices and encode the similarity between samples in their respective input space

KU LEUVEN

KERNEL FUSION SCHEMES



Given a training data ($x_1, x_2, \ldots, x_N$), a kernel matrix $K(x_i, x_j)$, for i,j = 1,2,…,N

$$\mathbf{K} = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \ldots & K(x_1,x_N) \\ K(x_2,x_1) & K(x_2,x_2) & \ldots & K(x_2,x_N) \\ . & & . & . \\ . & & . & . \\ K(x_N,x_1) & K(x_N,x_2) & \ldots & K(x_N,x_N) \end{bmatrix}$$

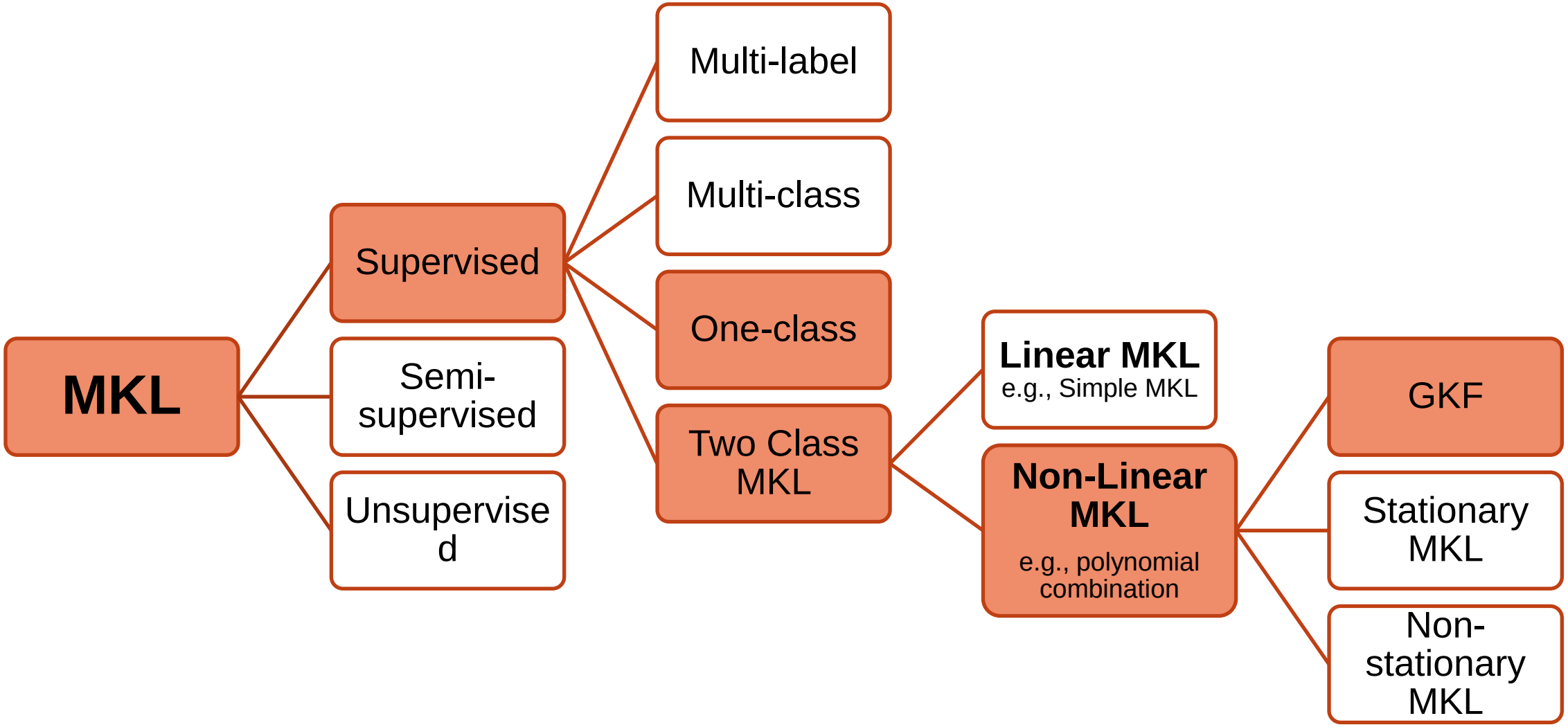
OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

Multiple kernel learning motivations

- Success of SVM is dependent on choice of good kernel:
  - Reducing bias due to kernel selection while allowing for more automated machine learning methods,
  - How to choose kernels
    - Kernel function
    - Parameters
- Most real problems involve multiple heterogeneous data sources
  - How can kernels help to fuse features from different data sources that) that have different notions of similarity
    - Particularly data sources that have different notions of similarity

Multiple Kernel Learning, tasks and strategies



Multiple kernel learning approaches

Fixed Rules	• Based on functions that do not require any training and have no parameters.
Heuristic Approaches	• Obtaining a weight of each kernel function independently
Similarity Optimization	• Maximizing the similarity to ideal kernel matrix
Structural Risk Optimization	• Minimizing "regularization term" + "error term"
Boosting approaches	• based on ensemble and boosting procedures, add a new kernel iteratively until the performance plateaus
Bayesian approaches	• Putting priors on the kernel parameters and learn the parameter values from the priors and the base algorithm

Multiple kernel learning

General MKL:

$$K_\beta(x_i, x_j) = f_\beta\left(\{K_m(x_i^m, x_j^m)\}_{m=1}^{P}\right)$$

Linear MKL:

$$K_\beta(x_i, x_j) = \sum_{m=1}^{P} \beta_m K_m(x_i^m, x_j^m)$$

Structure of optimization problem for MKL:

Introducing a new kernel K for n kernels β_i is a vector of coefficients for each kernel. $K = \sum_{m=1}^{p} \beta_m K_m$

For a set of data X with labels Y, he minimization problem can then be written as

$$\min_{\beta,c} E(Y, Kc) + R(K, c)$$

E is an error function

R is a regularization term

the square loss function or the hinge loss function

l_n norm or some combination of the norms

KU LEUVEN

MKL: Similarity Optimization

- Similarity:
  - kernel alignment
  - Euclidean distance
  - Kullback-Leibler (KL) divergence

$$A(\mathbf{K}_1, \mathbf{K}_2) = \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle \langle \mathbf{K}_2, \mathbf{K}_2 \rangle}}$$

$$A(\mathbf{K}, \mathbf{y}\mathbf{y}^T)$$

$$\langle \mathbf{K}_1, \mathbf{K}_2 \rangle = \sum_i \sum_j K_1(x_i^1, x_j^1) K_2(x_i^2, x_j^2)$$

MKL: Similarity Optimization

- Lanckriet et al. (2004)

$$\max_{\beta} A(\mathbf{K}_\beta, \mathbf{y}\mathbf{y}^T)$$
$$\text{s.t. } tr\left(\mathbf{K}_\beta\right) = 1, \qquad \mathbf{K}_\eta \succ= 0 \qquad \mathbf{K}_\beta = \sum_{m=1}^{P} \beta_m \mathbf{K}_m$$

Converted to a Semi-definite programming problem

Cortes et al (2010)

MKL: Structural Risk Optimization

$$\omega(\mathbf{K}_{\boldsymbol{\eta}}) = \max_{\boldsymbol{\alpha}} \mathbf{1}^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{Y} \mathbf{K}_{\boldsymbol{\eta}} \mathbf{Y} \boldsymbol{\alpha}$$

$$\text{s.t.} \quad \mathbf{1}^T \mathbf{Y} \boldsymbol{\alpha} = 0 \quad 0 \leq \boldsymbol{\alpha} \leq C$$

$$\Downarrow$$

$$\min_{\boldsymbol{\eta}} \omega(\mathbf{K}_{\boldsymbol{\eta}}) + r(\boldsymbol{\eta})$$

$$\text{s.t.} \quad \mathbf{K}_{\boldsymbol{\eta}} \succeq 0$$

Lanckriet et al. (2002)

MKL: Heuristic approaches

- Using a combination function that is parameterized. The parameters are generally defined for each individual kernel based on single-kernel performance or some computation from the kernel matrix.
- E.g., using a definition of kernel alignment

$$A(\mathbf{K}_1, \mathbf{K}_2) = \frac{\langle \mathbf{K}_1, \mathbf{K}_2 \rangle}{\sqrt{\langle \mathbf{K}_1, \mathbf{K}_1 \rangle \langle \mathbf{K}_2, \mathbf{K}_2 \rangle}}$$

- We can obtain a weight of each kernel as follow

Qui and Lane (2009)

$$\beta_m = \frac{A(\mathbf{K}_m, YY^T)}{\sum_{i=1}^{P} A(\mathbf{K}_i, YY^T)}$$

MKL: **Fixed rules approaches**

- Fixed rules approaches use rules to set the combination of the kernels.
- not require parameterization
- They use rules like summation and multiplication to combine the kernels

$$\mathrm{K} = 1/P \sum_{m=1}^{P} \mathbf{K}_m$$

- Gematric kernel fusion is a non-linear fixed rule-based approach

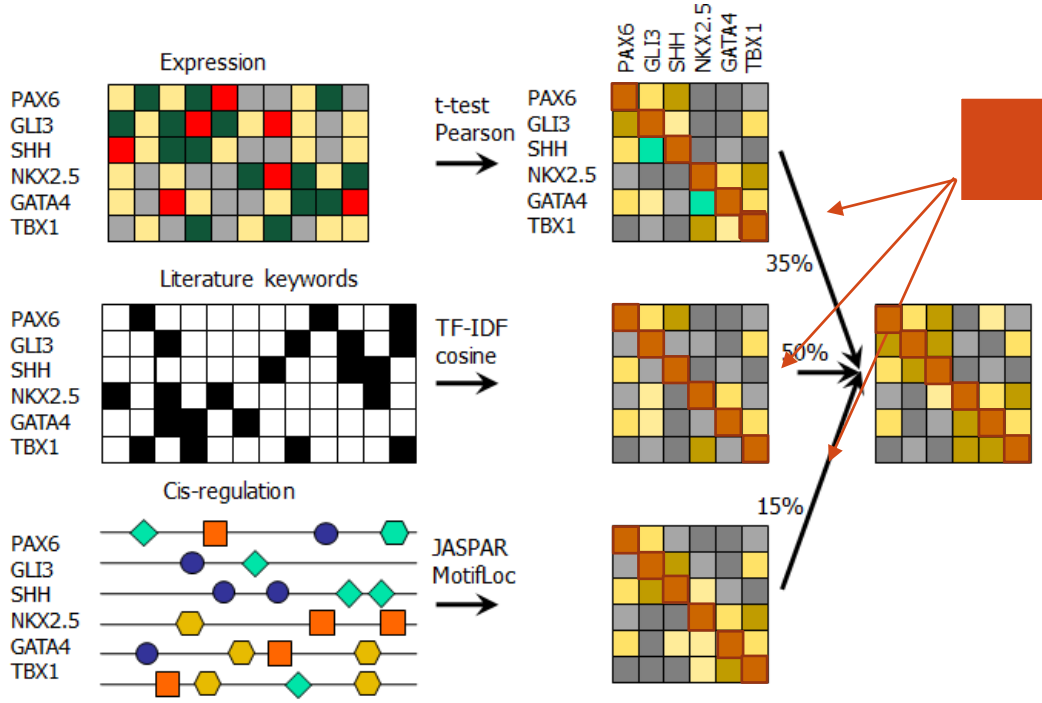
CHALLENGE OF HETEROGENEOUS DATA

HOW TO INTEGRATE HETEROGENEOUS BIOLOGICAL DATA IN A SUCCINCT AND INTUITIVE MANNER?

M views on data

- **Multiple Kernel Learning (MKL)**
  - **Combine heterogeneous sources of data in natural way**

$$K(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{M} d_m K^m(\mathbf{x},\mathbf{x}'), \text{ with } d_m \geq 0, \sum_m d_m = 1.$$



Optimizing kernel weights

When does the optimization of kernel weights really cause a significant improvement ?

KU LEUVEN

BAD NEWS ABOUT MKL

- Limited in that they focus on learning linear combination of base kernels.
  - Do not consider what are appropriate feature spaces for a given task.
  - Hiccup 1 going with such approaches could cause a loss of potentially useful latent information in biological data.
- It is often reduced to a convex optimization problem.
  - Solving this optimization problem is only possible for a small number of kernels and data points.
  - Hiccup 2 Biological data is often in large quantities .
- Only in case of redundant or too noisy data set, optimization of weights leads to a better performance.
  - Sensitive in dealing with complementary or noisy kernels.
  - Hiccup 3 Kernels in biological applications often encode the complementary characteristics of biological data.

Moreover, for a limited number of kernels, results obtained by employing the averaging of the kernels are comparable to the results of the best existing MKL approaches in general applications.

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

AVERAGE IS BEAUTIFUL

- **Wainer's theorem (1976):**
  Under very general circumstances, coefficients in multiple regression model can be replaced with equal weights with almost no loss in accuracy on the original sample.

  - All optimized weights are uniformly distributed on the interval [0.25; 0.75], use of equal weights barely changes performance.

- In biological data integration, all data sets are standardized, meaning that Wainer's theorem holds when no data set is completely redundant.
  - All data sets are relevant
  - No data set containing sufficient information for perfect job.

We would therefore opt for equal weights.

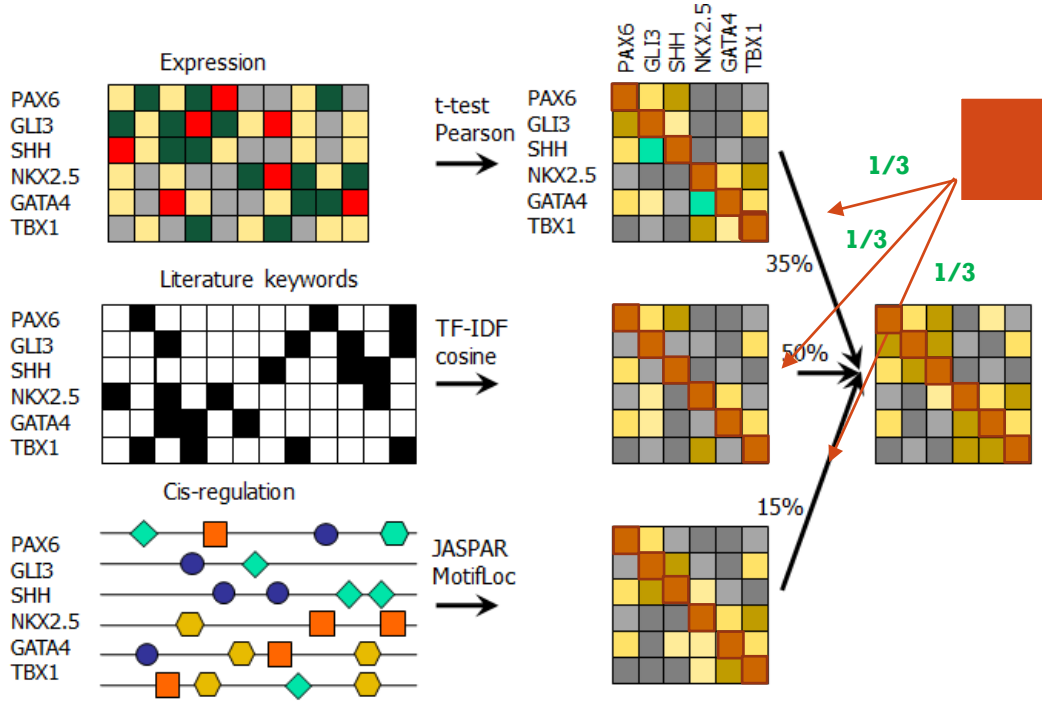
AVERAGE IS BEAUTIFUL, BUT…

HOW TO INTEGRATE HETEROGENEOUS BIOLOGICAL DATA IN A SUCCINCT AND INTUITIVE MANNER?

M views on data

- **Multiple Kernel Learning (MKL)**
  - **Combine heterogeneous sources of data in natural way**

$$K(\mathbf{x},\mathbf{x}') = \sum_{m=1}^{M} d_m K^m(\mathbf{x},\mathbf{x}'), \text{ with } d_m \geq 0, \sum_m d_m = 1.$$



Optimizing kernel weights

The linear-based averaging of kernel matrices leads to mixed results

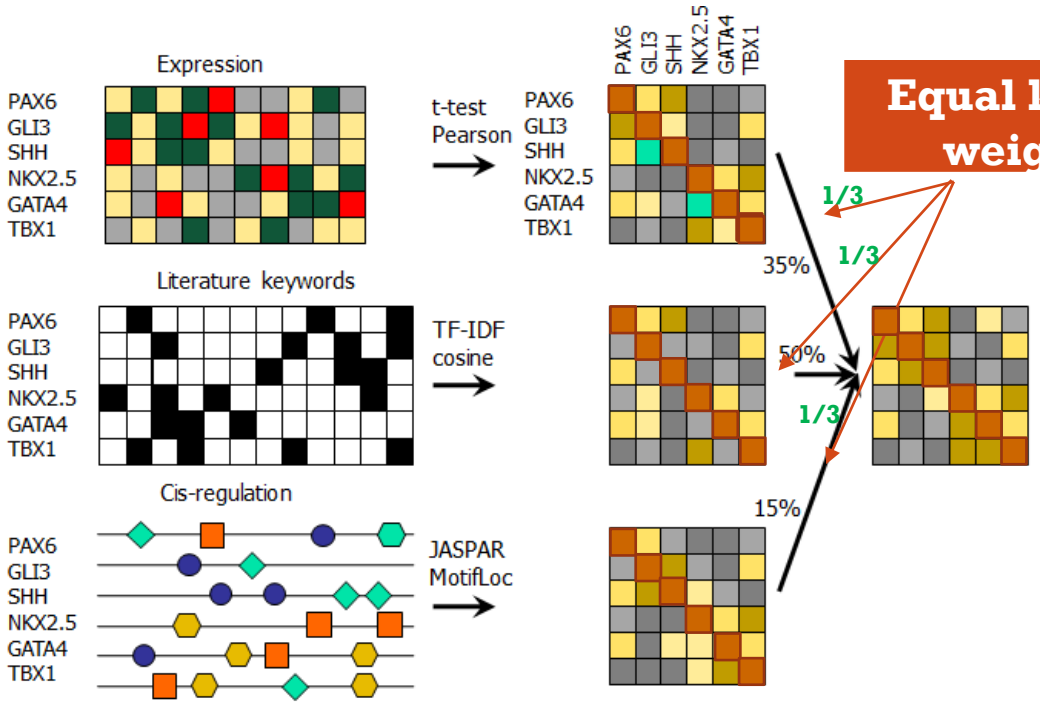
KU LEUVEN

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

CHALLENGE OF HETEROGENEOUS DATA

HOW TO INTEGRATE HETEROGENEOUS BIOLOGICAL DATA IN A SUCCINCT AND INTUITIVE MANNER?



Equal kernel weights

$$\mathcal{A}(A_1, \ldots, A_k) = \min_{X \in \mathbb{R}^{n \times n}} \sum_{i=1}^{k} \|X - A_i\|_F^2.$$

$$X = \sum_{i=1}^{k} \beta_i A_i \qquad \beta_1 = \beta_2 = \ldots = \beta_k = \frac{1}{k}$$

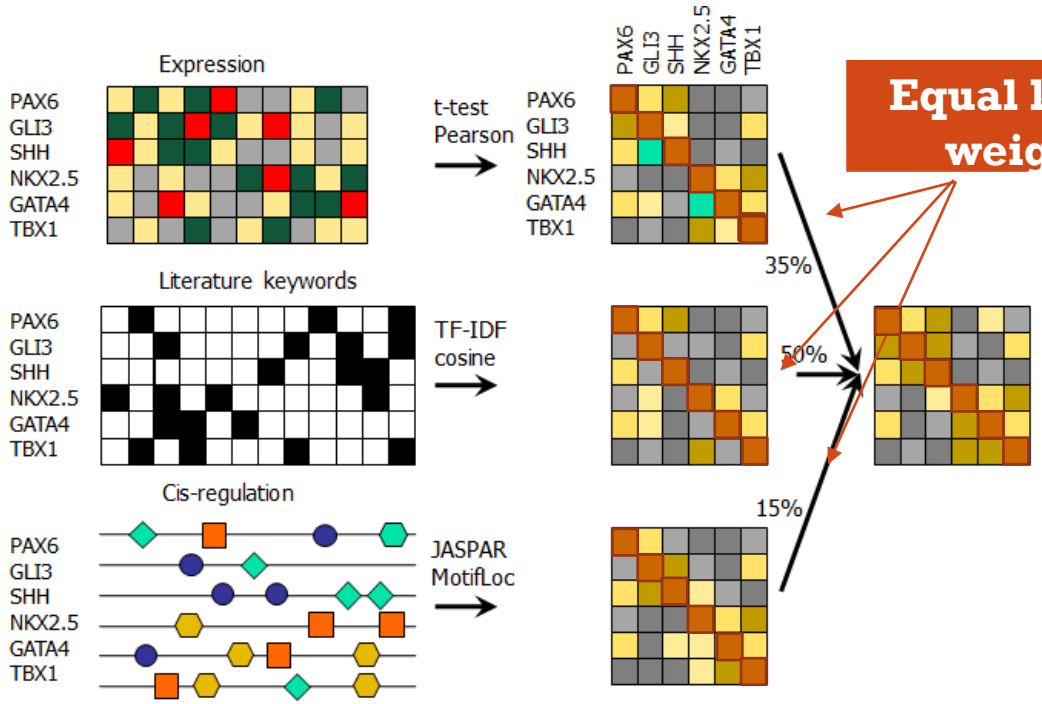
M views on data

SPD matrices form a convex cone and not a vector space.

Relying on concepts from Riemannian geometry

CHALLENGE OF HETEROGENEOUS DATA

HOW TO INTEGRATE HETEROGENEOUS BIOLOGICAL DATA IN A SUCCINCT AND INTUITIVE MANNER?



Equal kernel weights

$$\mathcal{A}(A_1, \ldots, A_k) = \min_{X \in \mathbb{R}^{n \times n}} \sum_{i=1}^{k} \|X - A_i\|_F^2.$$

$$X = \sum_{i=1}^{k} \beta_i A_i \qquad \beta_1 = \beta_2 = \ldots = \beta_k = \frac{1}{k}$$

M views on data

SPD matrices form a convex cone and not a vector space.

Relying on concepts from Riemannian geometry

Karcher mean $O(kn^3)$

Arithmetic-Geometric-Harmonic (AGH) mean

$O(n^2 \log(n) k)$

A descent approximation to the Karcher mean

$$\mathcal{G}(A_1, \ldots, A_k) = \min_{X \in \mathcal{P}_n} \sum_{i=1}^{k} \|\log(A_i^{-1/2} X A_i^{-1/2})\|_F^2,$$

GEOMETRIC KERNEL FUSION, GEOMETRIC MEAN

- For a general number of matrices, the fused kernel is obtained by taking the Geometric Mean

$$K = F(K_1, K_2, \ldots, Kn) = G(K_1, K_2, \ldots, Kn)$$

- The geometric mean of two PD matrices A and B can be defined explicitly as

$$G(A,B) = A(A^{-1}B)^{1/2} = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}.$$

However, for more than two matrices a proper definition remained elusive for a long

KARCHER MEAN

- The Karcher mean of SPD matrices $A_1, \ldots, A_k$ is defined as the barycenter of these matrices on the manifold of SPD matrices with its Riemannian geometry.
- This is obtained by searching the minimizer of an optimization problem, given as follows:

$$\mathcal{G}(A_1, \ldots, A_k) = \min_{X \in \mathcal{P}_n} \sum_{i=1}^{k} || \log(A_i^{-1/2} X A_i^{-1/2}) ||_F^2,$$

Algorithm 1 The Karcher mean optimization algorithm

Let $A_1, \ldots, A_k$ be SPD matrices, X_0 an initial guess, and R_X a retraction starting at a point X

- for $k = 0, 1, \ldots$
  - Determine the search direction ξ_k using Steepest Descent, Conjugate Gradient, Newton, ...
  - $X_{k+1} = R_{X_k}(t^A \xi_k)$ where t^A is an appropriate stepsize
- end

ARITHMETIC-GEOMETRIC-HARMONIC (AGH) MEAN

A descent approximation to the Karcher mean.

Computationally scalable approach for geometric mean Using AGH mean

For each iteration:

$O(n^2 \log(n) k)$ vs $O(kn^3)$ Karcher mean

Algorithm 1 The approximate AGH mean algorithm where $\mathcal{A}$ denotes the AM and $\mathcal{H}$ the HM

Let $A_1, \ldots, A_k$ be SPD matrices

- For all i set $B_i = A_i$ and $C_i = A_i$;
- while not converged
  - For all i set $\tilde{B}_i = \mathcal{H}(B_i, C_{(i \bmod n)+1})$;
  - For all i set $\tilde{C}_i = \mathcal{A}(B_i, C_{(i \bmod n)+1})$;
  - For all i set $C_{p(i)} = \tilde{C}_i$, $B_{p(i)} = \tilde{B}_i$, with p a random permutation of $[1, \ldots, n]$.
- end

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - **Scalable methods for geometric kernel data fusion**
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

FROM GEOMETRIC KERNEL FUSION TO LOG-EUCLIDEAN MEAN

- If $x_1, x_2, \ldots, x_n$ are n positive numbers, then their geometric mean is given by:

$$G(x_1, \ldots, x_n) = (x_1, \ldots, x_n)^{\frac{1}{n}} = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \log(x_i)\right)$$

- A generalization of the notion of geometric mean.

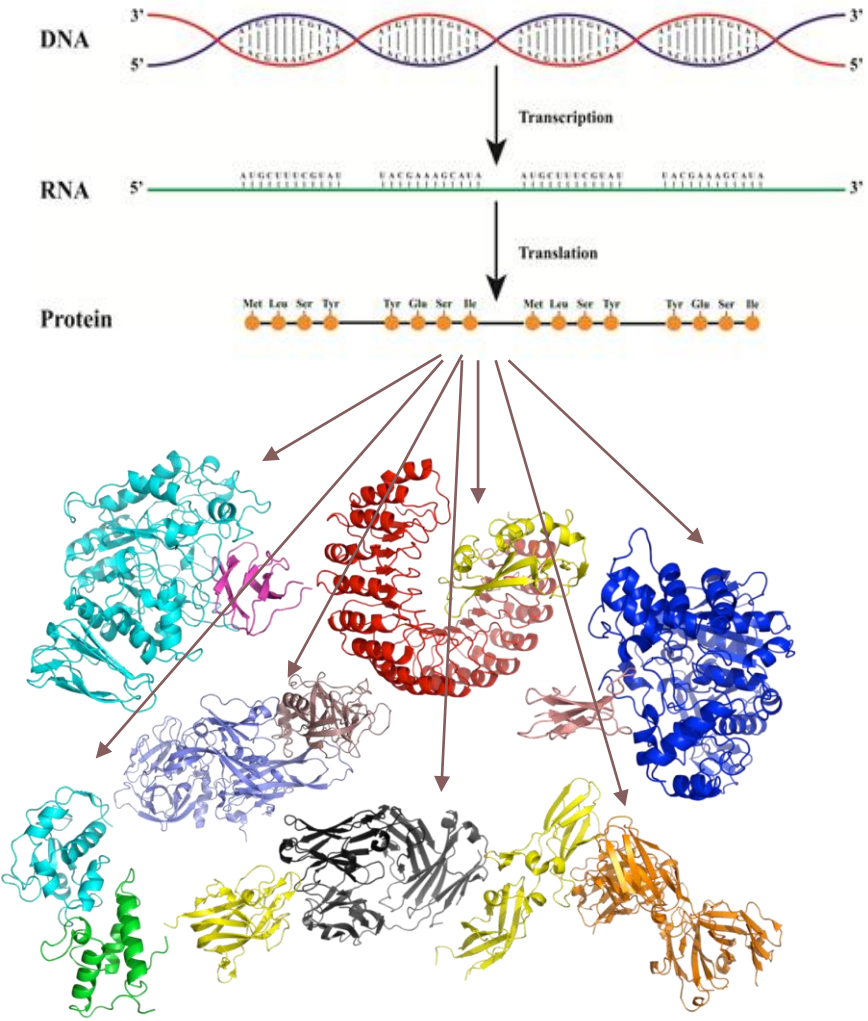
$$K_{LE}(K_1, \ldots, K_n) = \exp\left(\frac{1}{n} \sum_{i=1}^{n} \log(K_i)\right)$$

- Thanks to an one-to-one relation between the (Euclidean) vector space of symmetric matrices and the positive definite matrices using the matrix exponential and logarithm

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

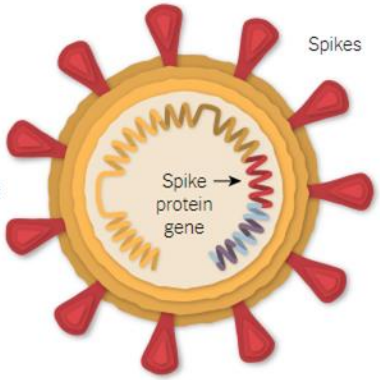
PROTEIN FOLD RECOGNITION

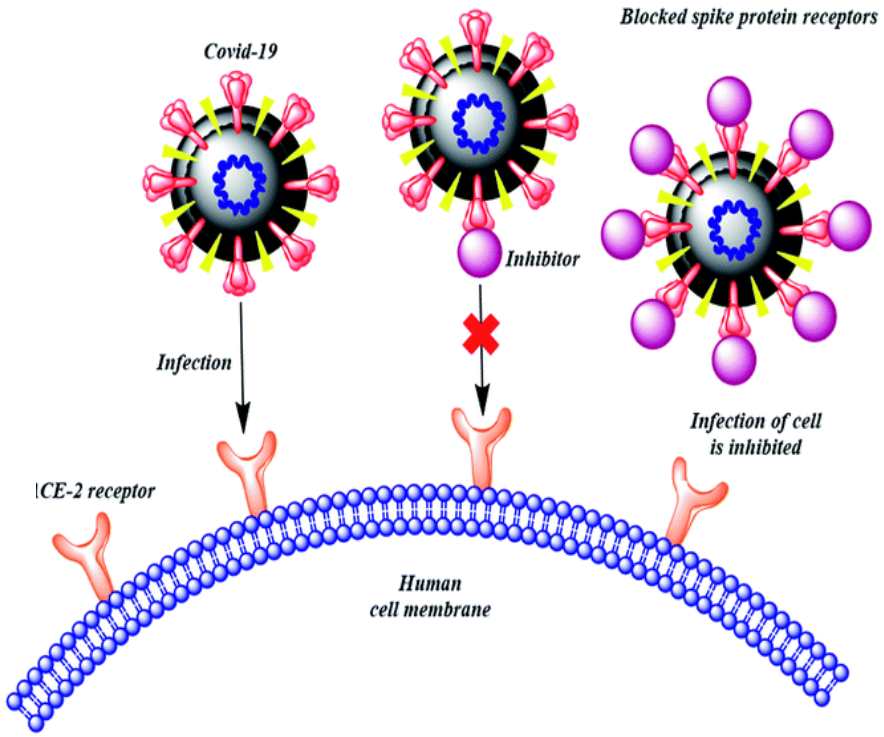


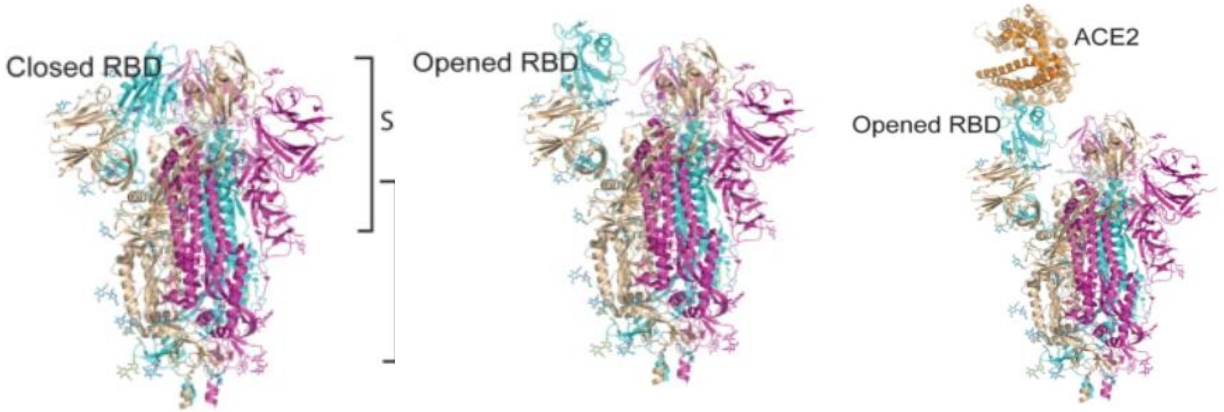
Sputnik V
AstraZeneca
Johnson & Johnson

Moderna
Pfizer–BioNTech

Novavax



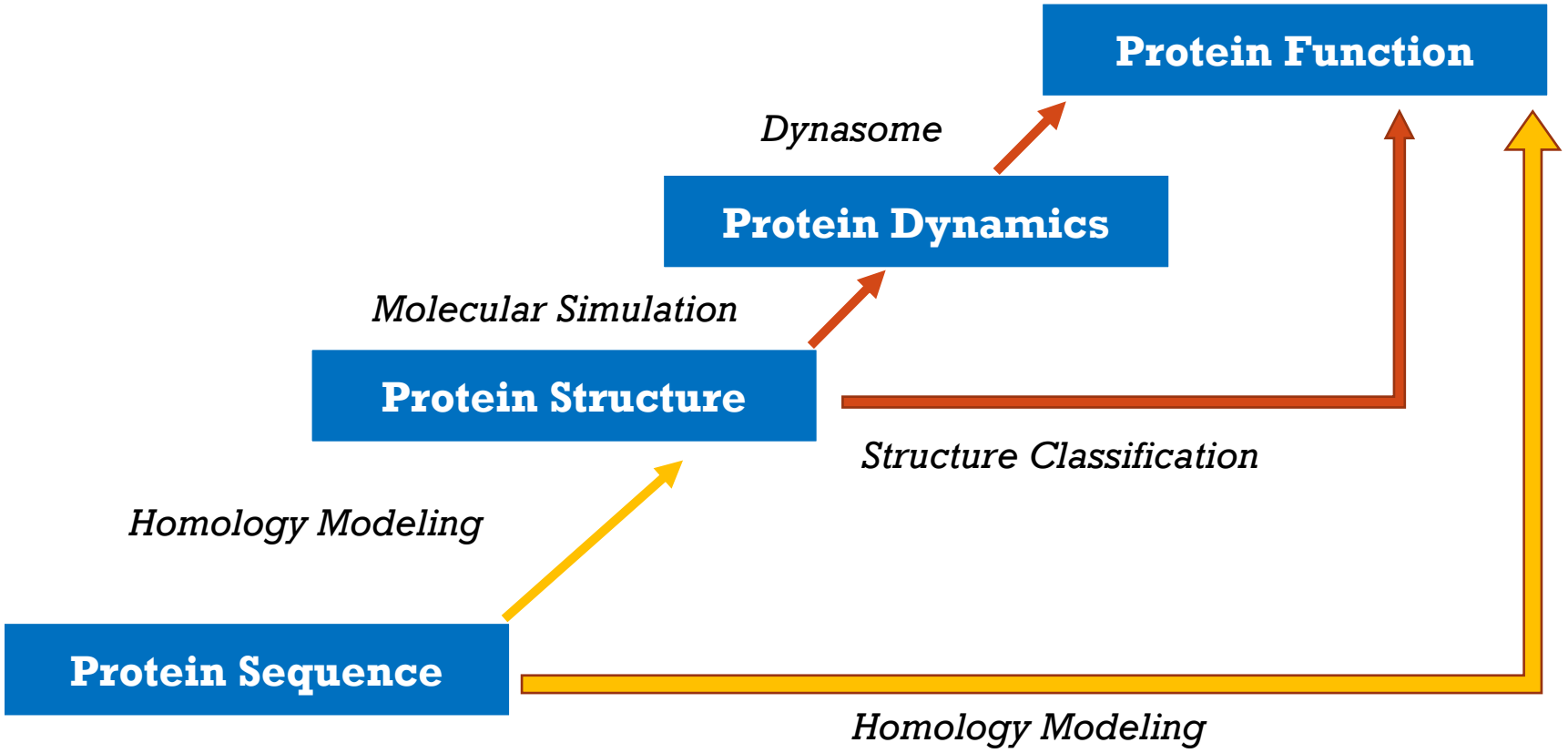




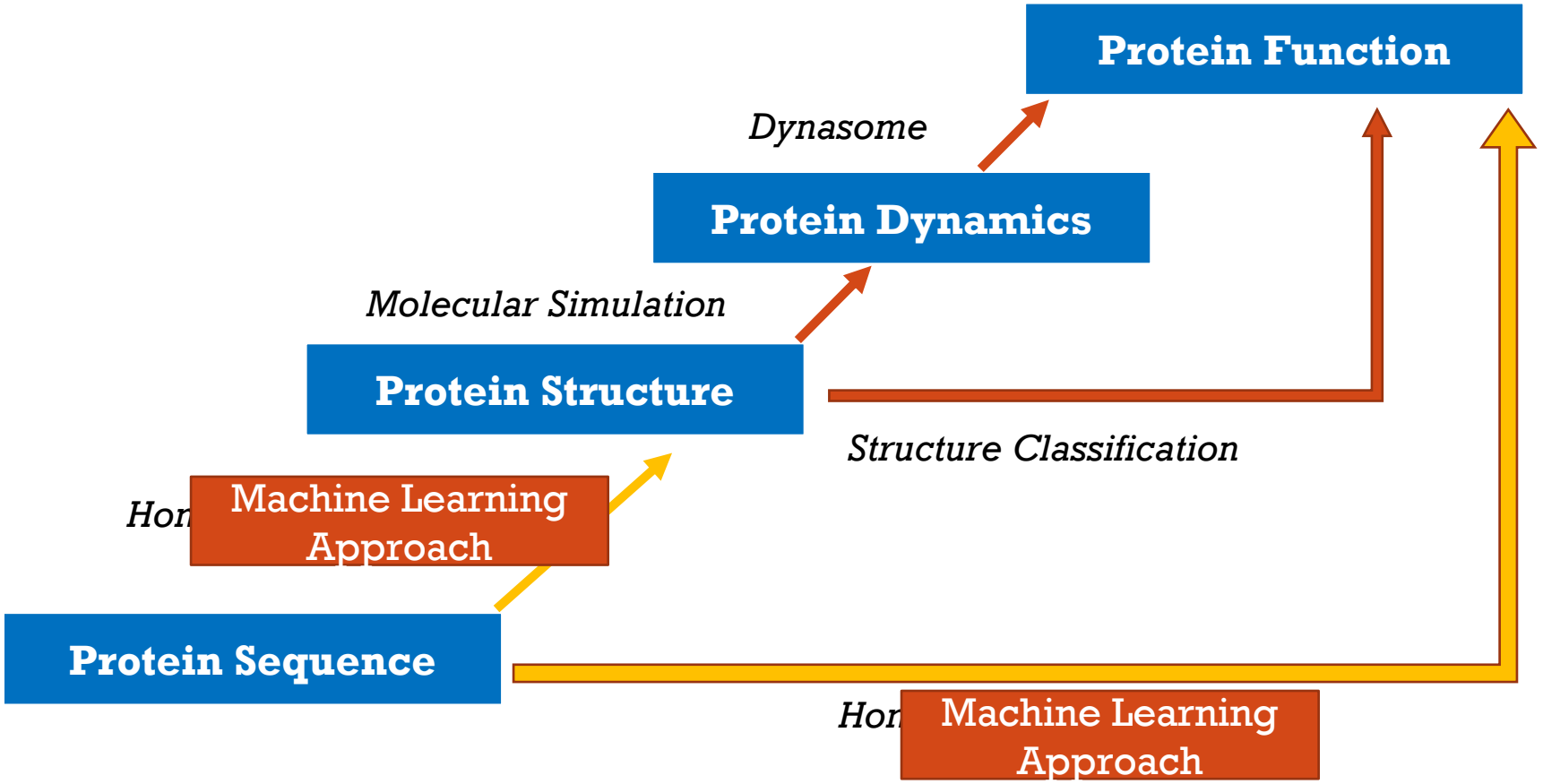
Huang, Y, *et al.* Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin* **41,** 1141–1149 (2020).

KU LEUVEN

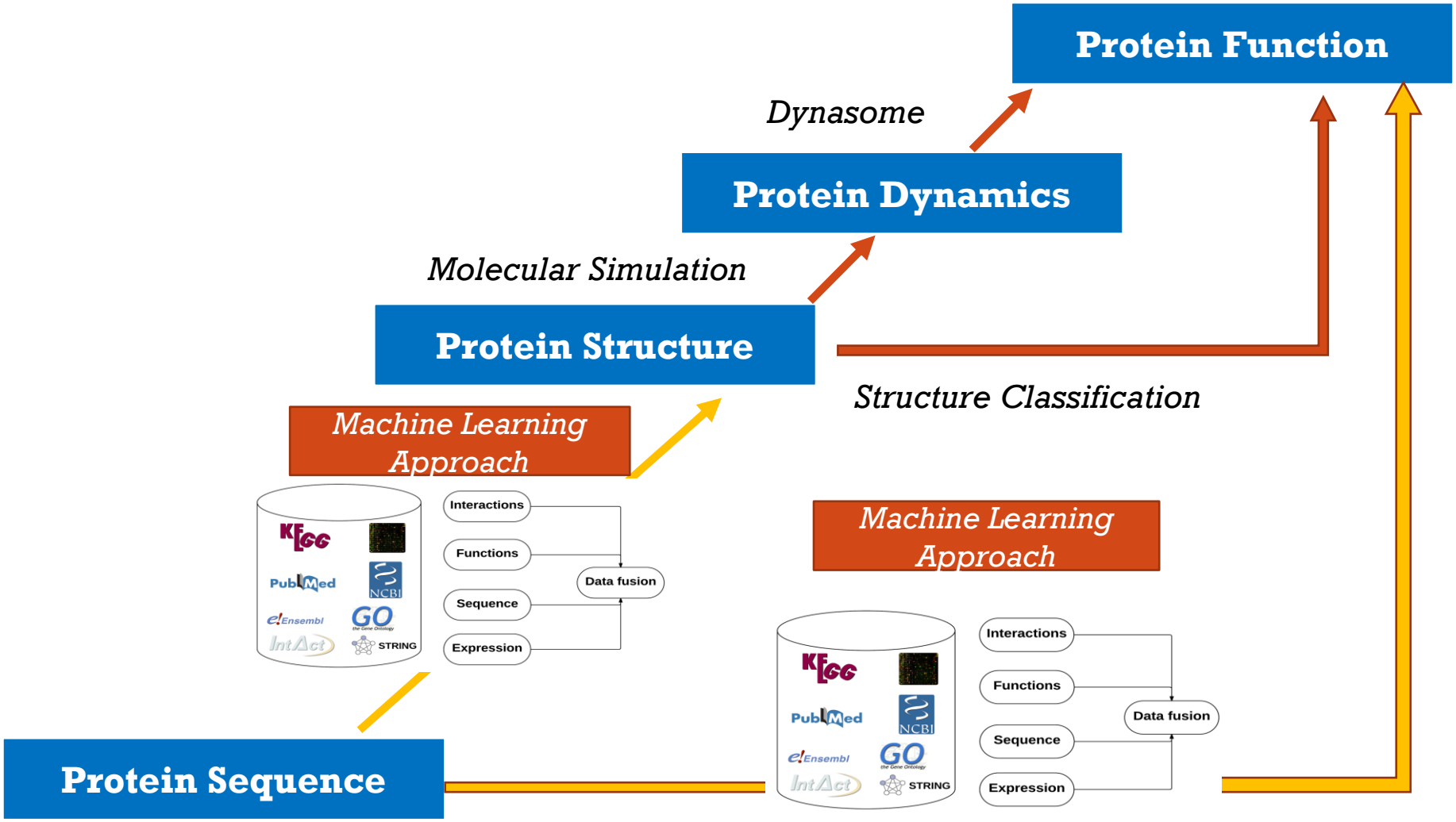
GOAL: FROM SEQUENCE INFORMATION TO PROTEIN STRUCTURAL AND FUNCTIONAL INFORMATION



GOAL: FROM SEQUENCE INFORMATION TO PROTEIN STRUCTURAL AND FUNCTIONAL INFORMATION



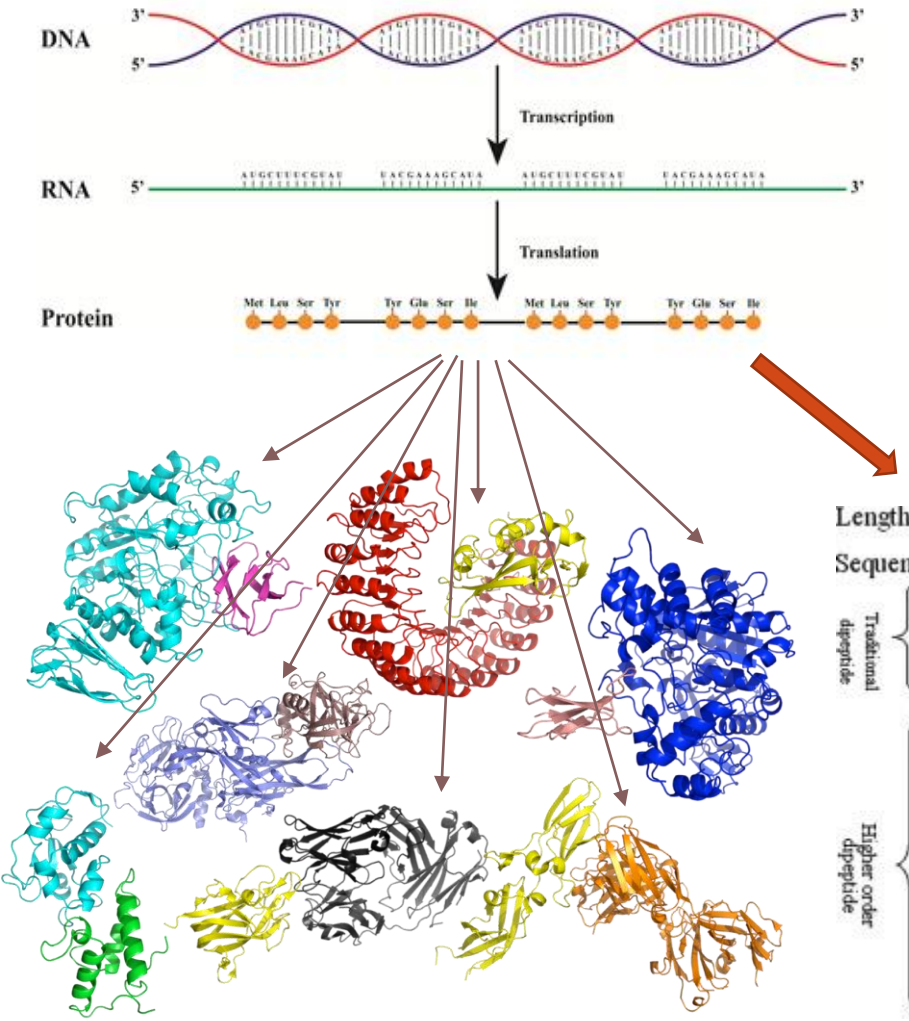
GOAL: FROM SEQUENCE INFORMATION TO PROTEIN STRUCTURAL AND FUNCTIONAL INFORMATION

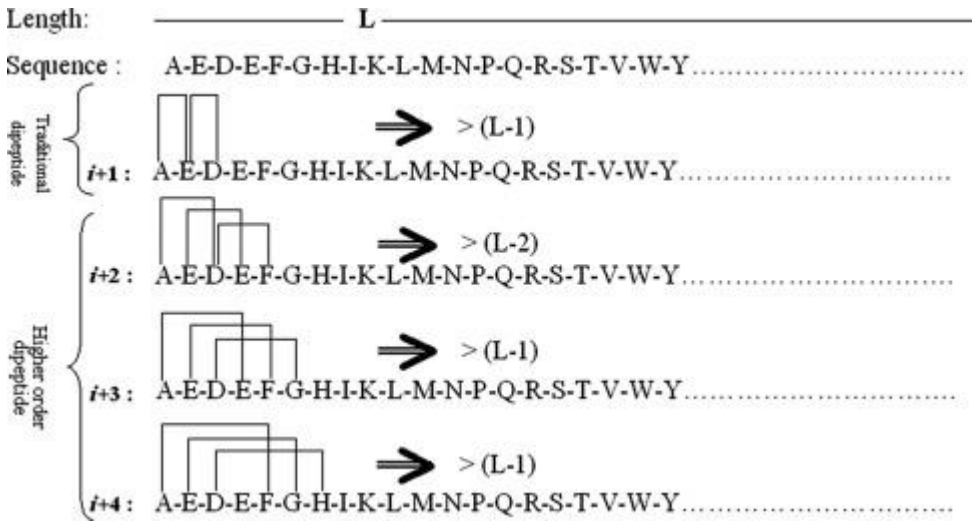


KU LEUVEN

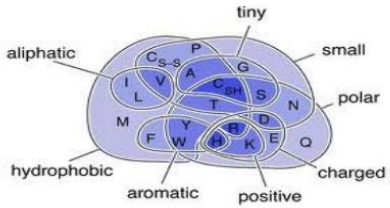
PROTEIN FOLD RECOGNITION

- Feature Vectors
  - 26 sequence-based protein features including
    - 6 types of structural information
      - Amino Acid composition (C),
      - Predicted Secondary Structure (S),
      - 4 PseAAC Compositions
    - 4 kinds of physicochemical properties
    - and 2 local pairwise sequence alignment-based feature spaces
    - sequence evolution information extracted directly from PSSM in 14 different ways.



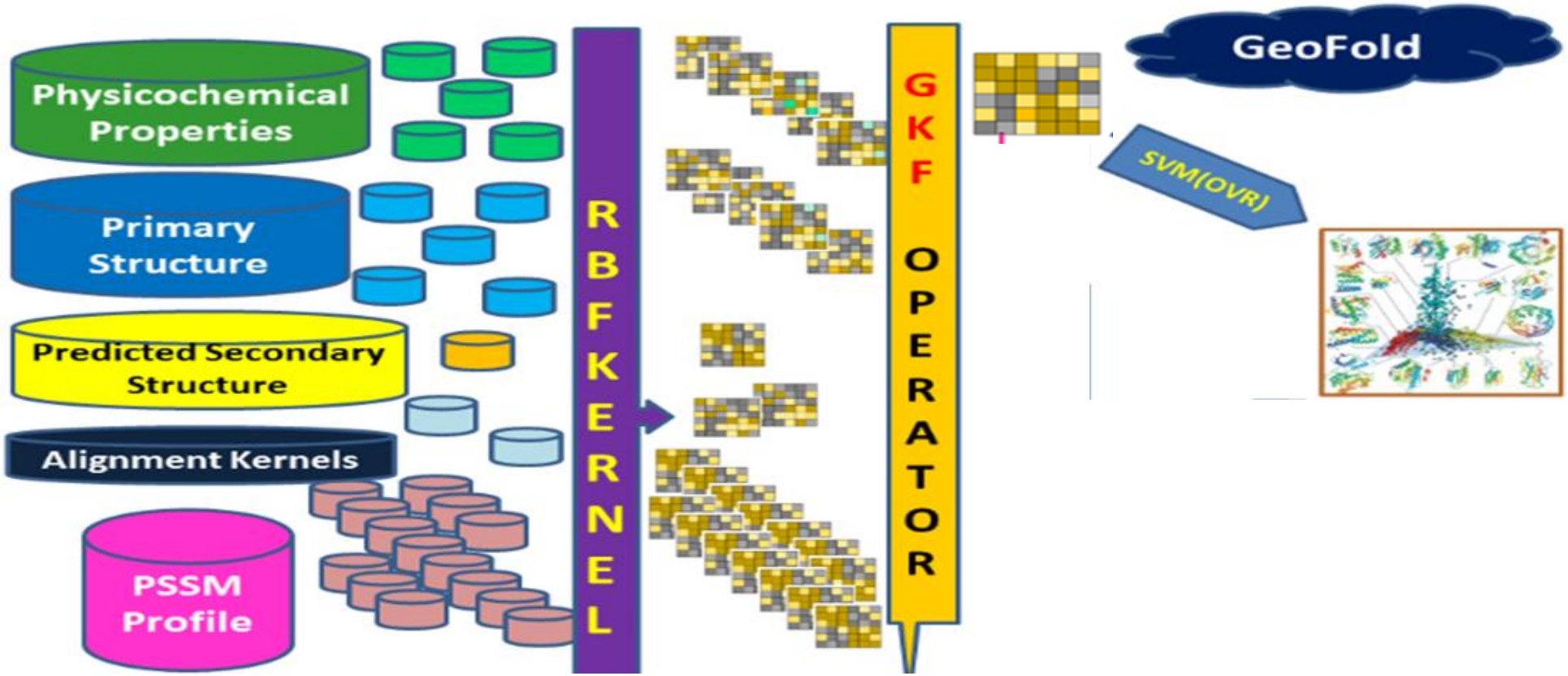


Characteristics and Properties of Amino Acids (AA)



PROTEIN FOLD RECOGNITION

- 26 sequence-based protein features

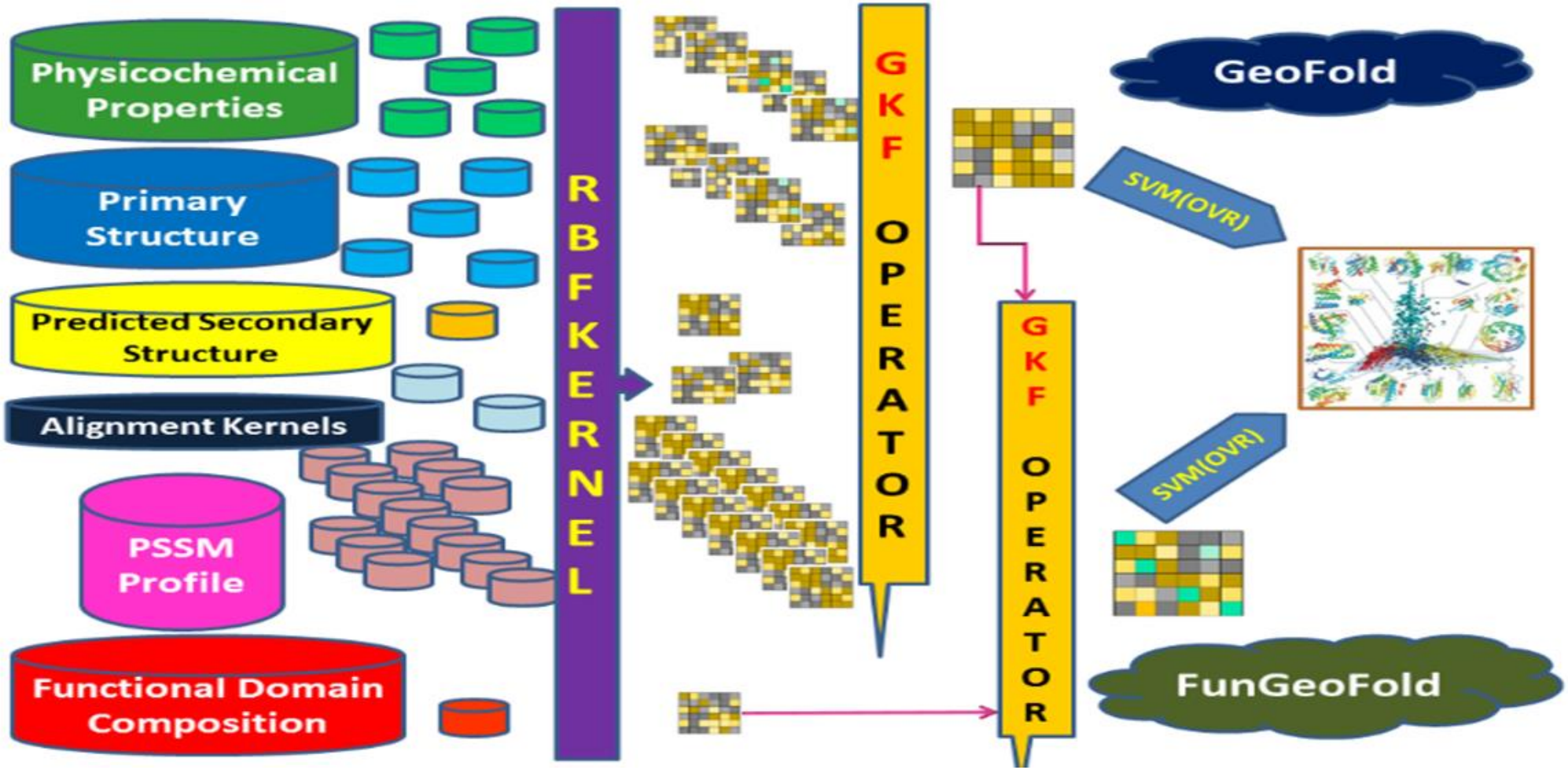


Zakeri P *et al*. “Protein fold recognition using geometric kernel data fusion,” Bioinformatics.,2014

KU LEUVEN

PROTEIN FOLD RECOGNITION (Incorporating the functional domain composition)

- 26 sequence-based protein features
- Incorporating the functional domain composition

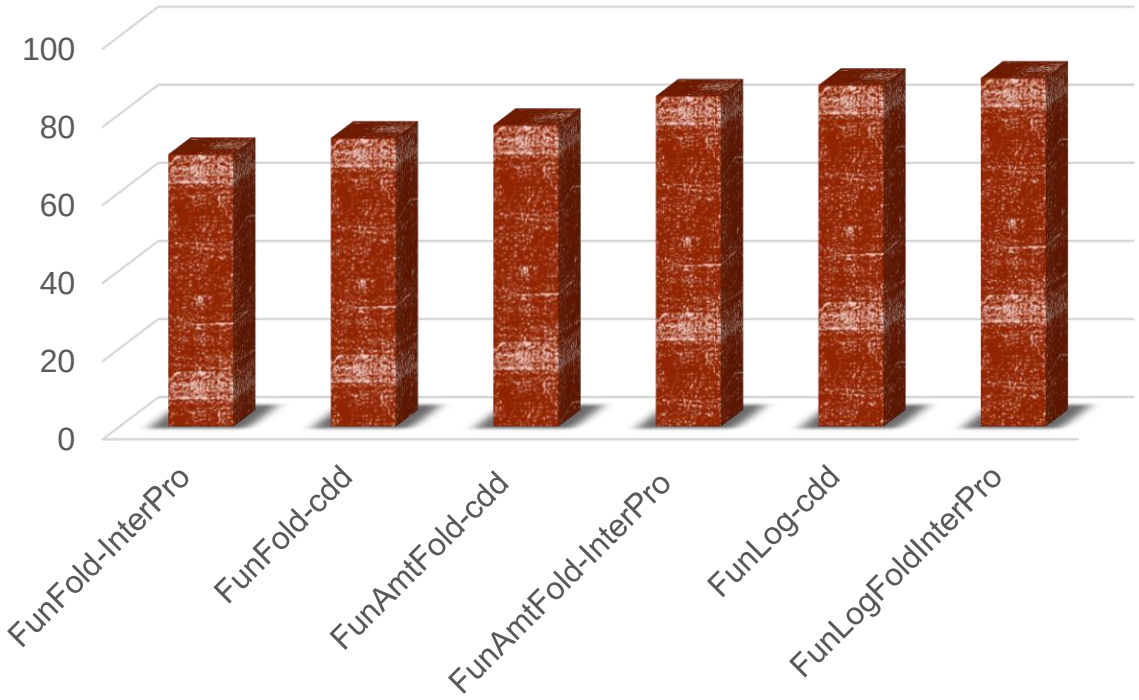


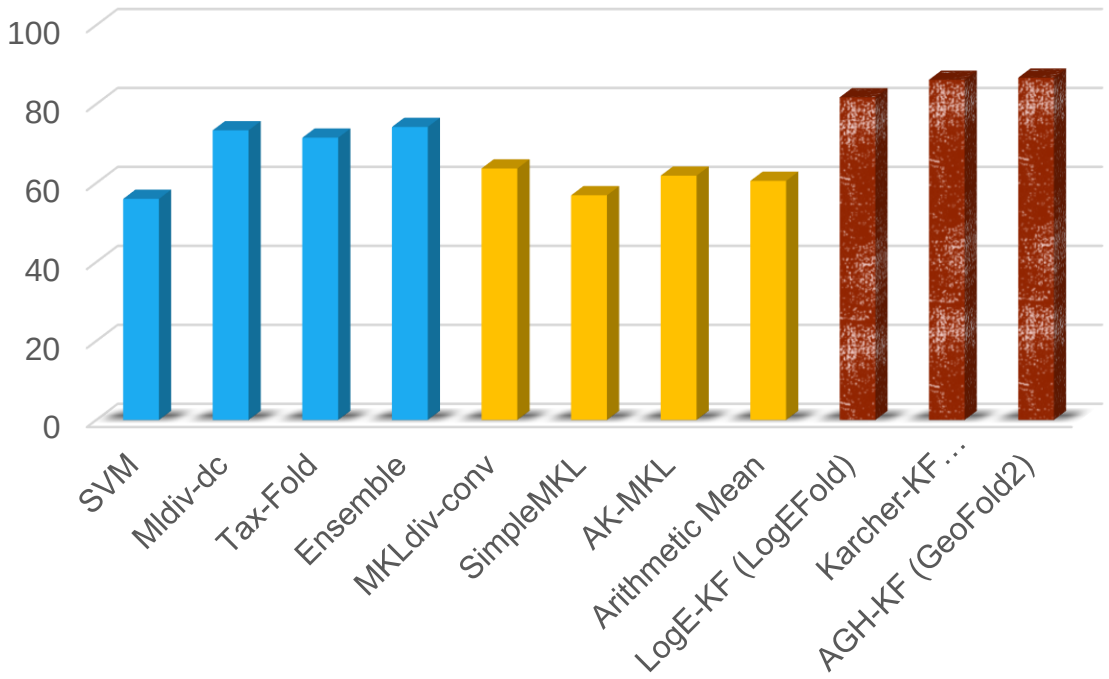
Zakeri P *et al*. “Protein fold recognition using geometric kernel data fusion,” Bioinformatics.,2014

KU LEUVEN

COMPARISON OF THE TOTAL ACCURACIES

- 26 sequence-based PFs
- Evaluated on DD data set (27 folds)
- Incorporating the functional domain composition

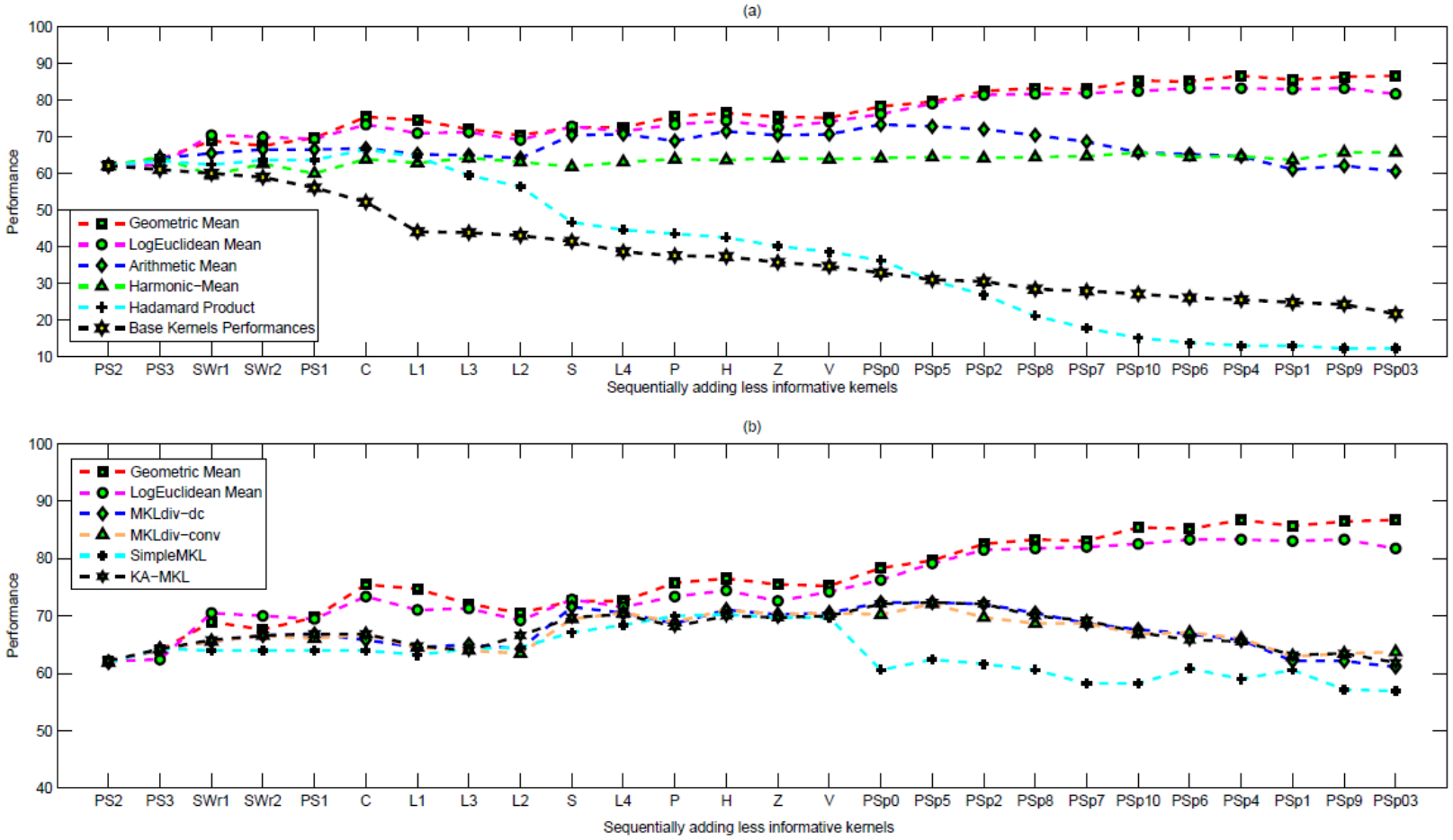




Zakeri P *et al*. “Protein fold recognition using geometric kernel data fusion,” Bioinformatics.,2014

KU LEUVEN

THE EFFECT OF SEQUENTIALLY INCORPORATING PROTEIN FEATURES



Zakeri P *et al*. “Protein fold recognition using geometric kernel data fusion,” Bioinformatics.,2014

KU LEUVEN

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

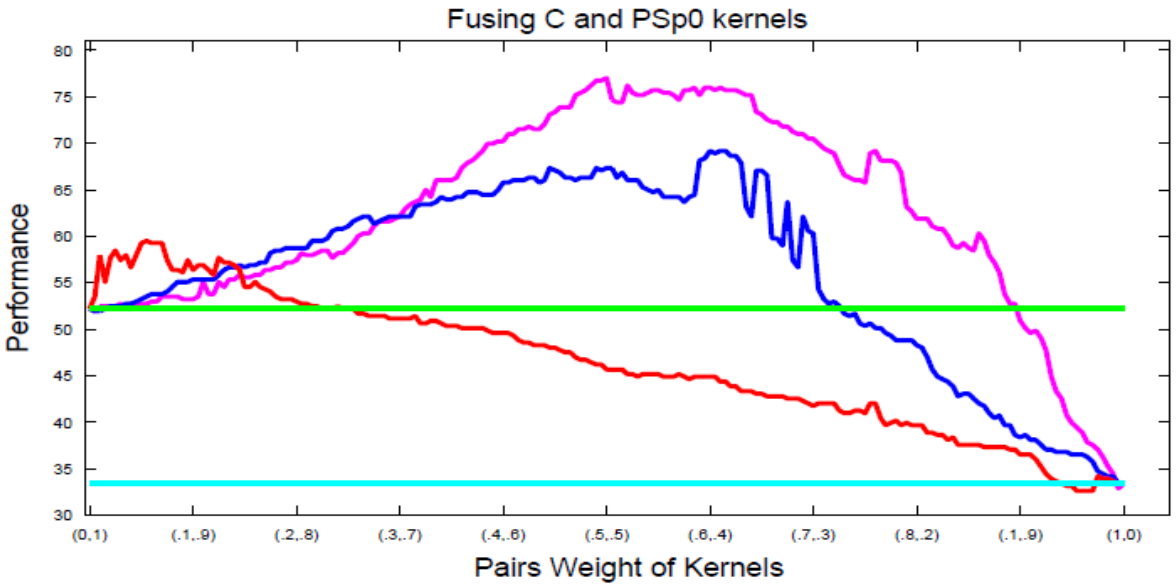
LIMITATION OF LINEAR COMBINATION

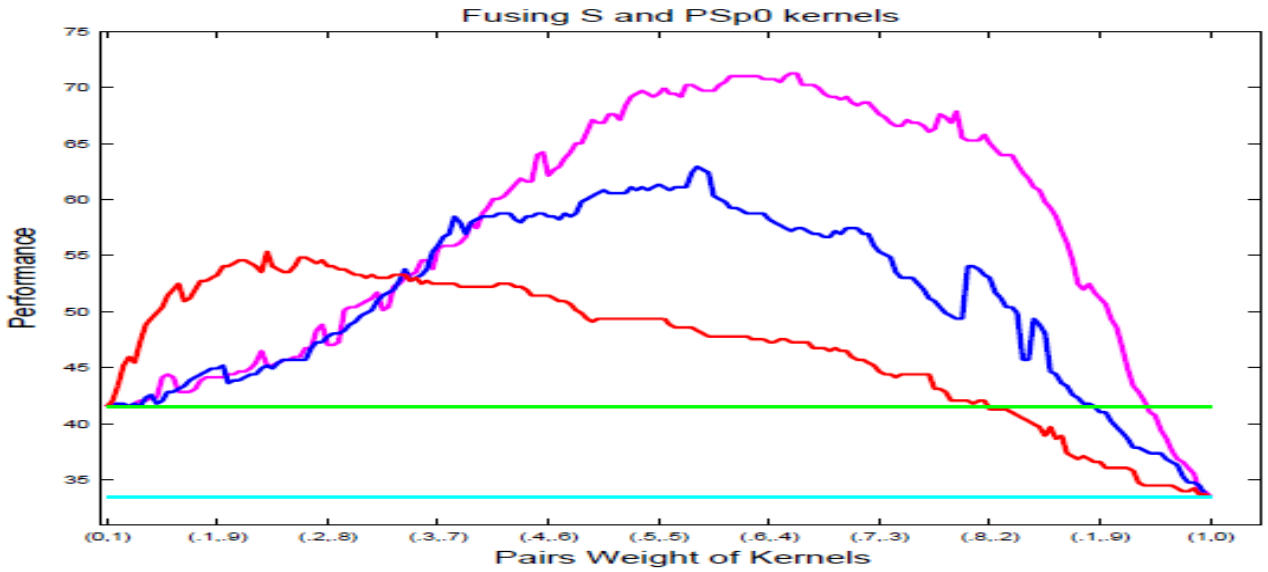
- 201 various convex combinations of two different kernels. Assign different weights to each kernel as follows:

$$K^i = w_i K_1 + (1 - w_i) K_2 \qquad 1 < i \leq 201$$

where w = [1, 0.995, 0.99, . . . , 0].

- Better success rates on the majority of the interval of kernel weight pairs for GKF



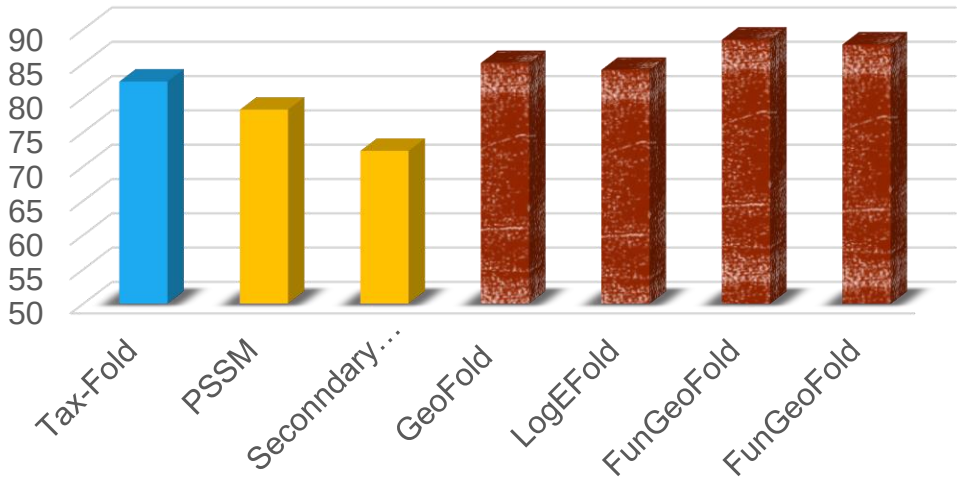


Zakeri P *et al*. “Protein fold recognition using geometric kernel data fusion,” Bioinformatics.,2014

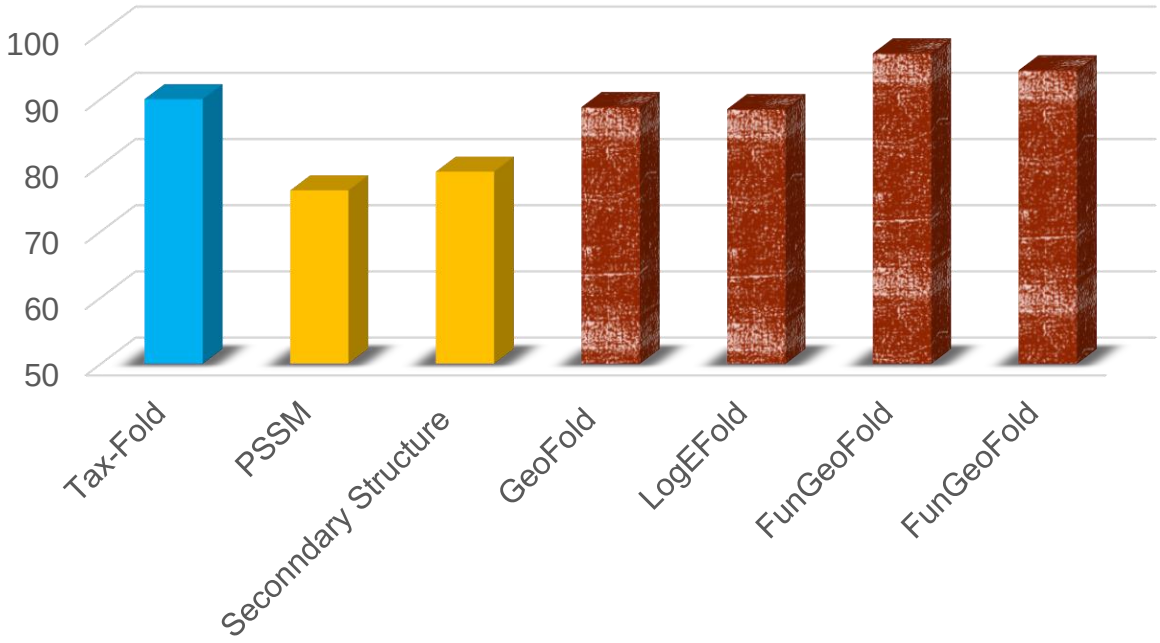
KU LEUVEN

RELATIONSHIP BETWEEN PRIMARY AND TERTIARY STRUCTURE

- ***Our results suggest that combining the evolutionary and secondary structural information could be crucial to elucidate such a latent link***
- 3397 protein sequences in 27 folds
- 6364 protein sequence in 95 folds



(95 folds)



(27 folds)

KU LEUVEN

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

CANDIDATE GENE PRIORITIZATION

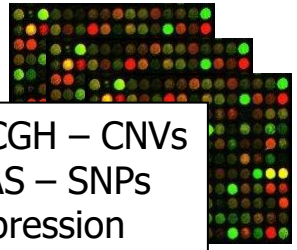
COVID GENE PRIORITIZATION

Training gene seed
Key Genes related to COVID19:
ACE2, TMPRSS2
Potential Genes related to COVID19:
IFNAR2, TYK2, OAS1, DPP9, CCR2

ML Model

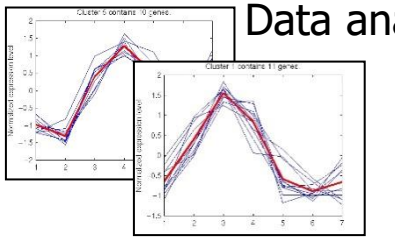
What about other ~ 20000 genes
APOE? ACE?

High-throughput genomics



Array CGH – CNVs
GWAS – SNPs
Expression
…

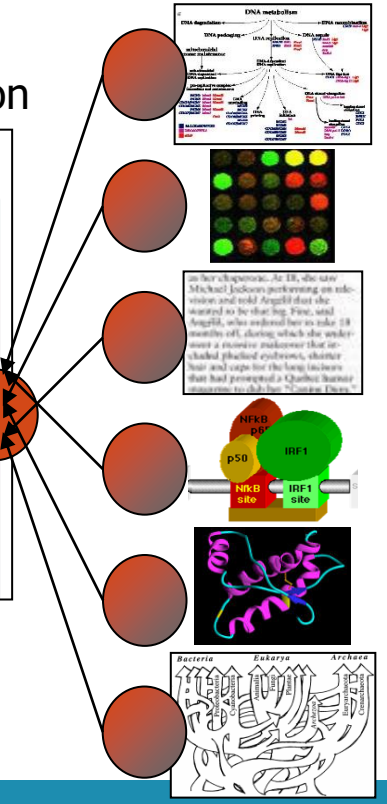
Data analysis



Candidate genes

Name	Ensembl
TTR	ENSG00000118271
PAH	ENSG00000171759
G6PC	ENSG00000131482
IGF1	ENSG00000017427
ALB	ENSG00000163631
CRP	00132693
HABP2	00148702
IF	00138799
FST	00134363
ARAF1	00000078061
HMGA2	G00000149948
C9	ENSG00000113600
PCBP2	G00000111406
HOXB6	G00000108511
RERE	ENSG00000142599
HOXA11	ENSG00000005073
CLIC1	ENSG00000096238
ERCC3	ENSG00000163161
ERCC3	ENSG00000163161
TLL2	ENSG00000095587
SYT4	ENSG00000132872
SYT4	ENSG00000132872
PIK4CB	ENSG00000143393
PKD2	ENSG00000118762
	ENSG00000081026
ANKRD3	ENSG00000183421
F13A1	ENSG00000124491
BPAG1	ENSG00000151914
KCNN3	ENSG00000143603
GRIN2A|GRIN2B	ENSG00000150086
SIM1	ENSG00000112246
	00000174891
	00000089195
C14orf10	00000092020
STX8	0000170310
	G00000107671
MSH5	000096474
CRH	000147571
MID1	000101871
	000184508
	00113460
TGFB3	ENSG00000
C1QR1	ENS
NR4	ENS
PDGF	ENSG0
PDGF	EN
NR3	EN
NFYA	
CP	ENS 7
TM	ENSG00 6537
MMP3|MM	ENSG00000149968

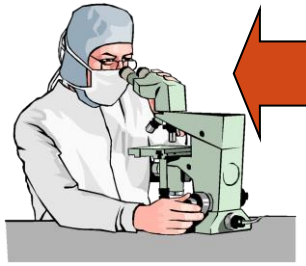
Information sources



Candidate prioritization

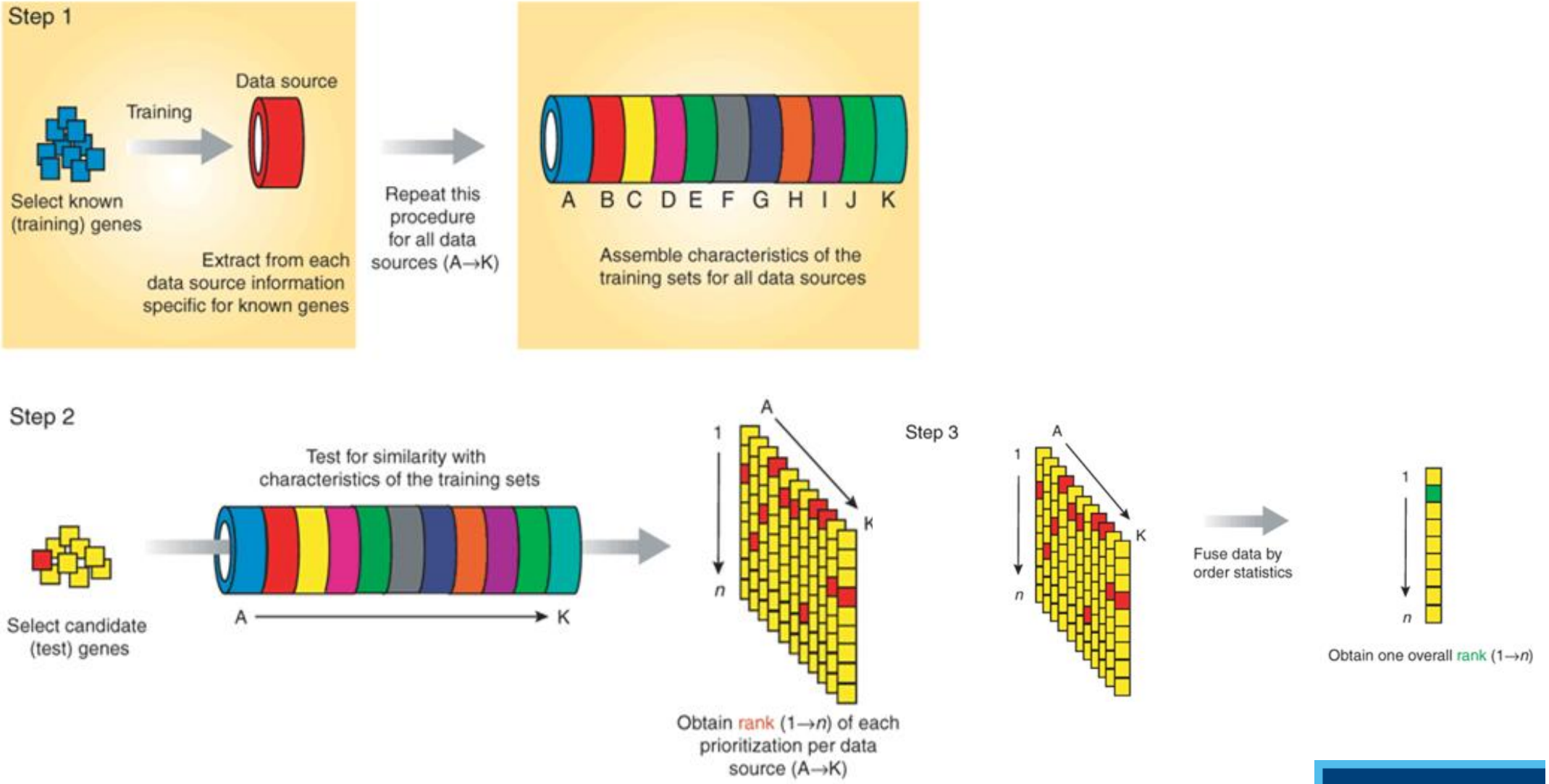
Rank	En	Ex	Ip	Ke	GO	Te	Avg	Pval
1	TTR	G6PC	PAH	G6PC	IGF1	TTR		TTR
2	IGF1	TTR	IGF1	PAH	PAH	IGF1		PAH
3	CRP	ALB	TTR	RERE	G6PC	CRP		G6PC
4	HOXB6	HABP2	ALB	ERCC3	TTR	HOXB6		IGF1
5	ALB	PAH	HDC	ERCC3		ALB		
6	NR4A2	IF	TLL2	ANKRD3		HMGA2		
7	PAH		C1QR1	ARAF1	HDC	NR4A2		
8	HOXA11	IGF1	G6PC	PKD2	F13A1	PAH		IF
9	NFYA	CRP	HABP2	MTMR1	KCNN3	HOXA11	C13orf7	FST
10	C9	ARAF1	IF	HDC	CLIC1	NFYA	TTR	ARAF1

Validation



KU LEUVEN

DISEASE GENE PRIORITIZATION using Data fusion (ENDEAVOUR)



S. Aerts *et al.*, "Gene prioritization through genomic data fusion". Nat Biotech, 2006

KU LEUVEN

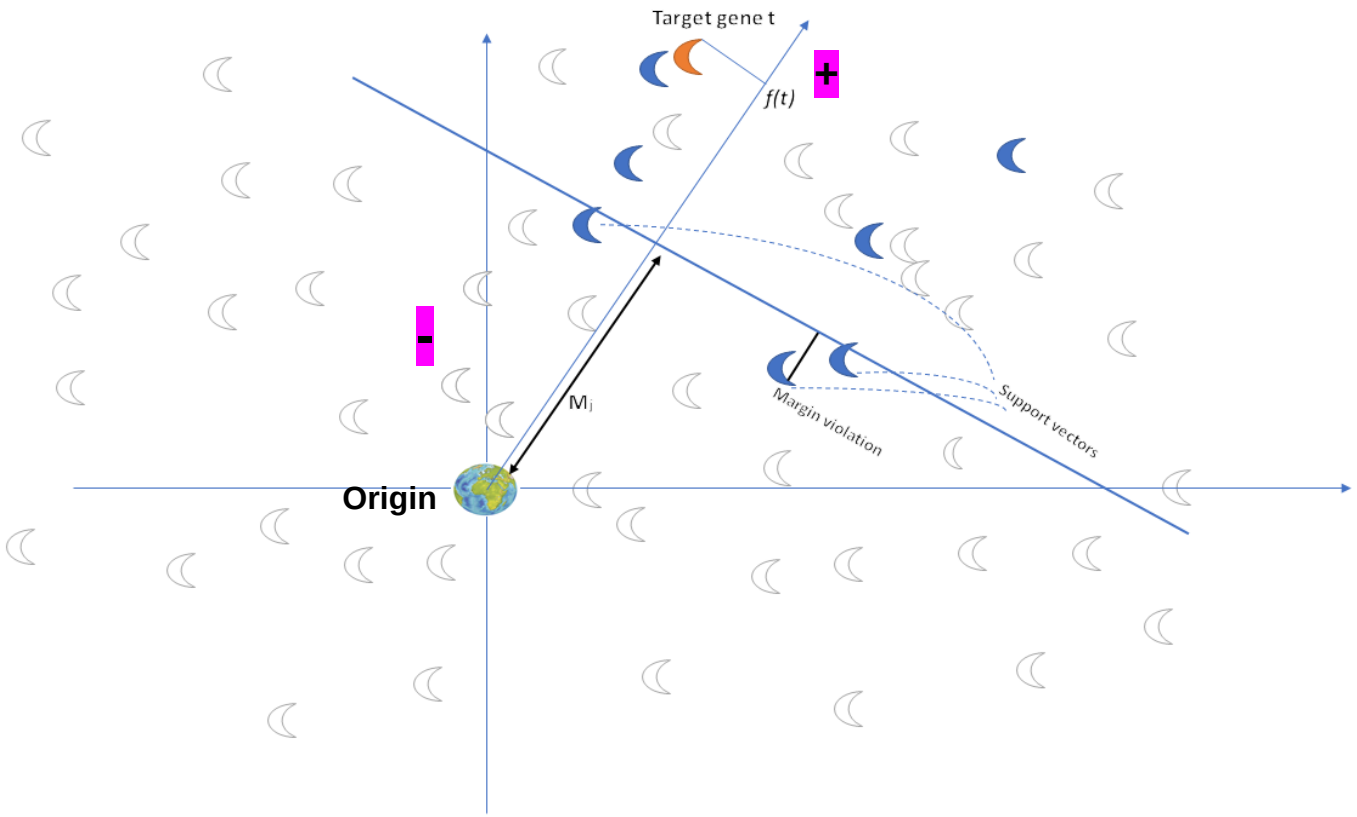
OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

One Class Support Vector Machine (OCSVM) Algorithm

- The one-class SVM is first discussed by Schölkopf
- Maps input data into a high dimensional feature space
- Iteratively finds the maximal margin in the hyperplane which best separates the training data from the origin
- Solves optimization problem to find rule f with maximal margin
  - $f(\mathbf{x})=\langle \mathbf{w},\mathbf{x}\rangle+b$
  - If $f(\mathbf{x})<0$, label $\mathbf{x}$ as anomalous

OCSVM


Target gene t
f(t)
+
-
M_i
Margin violation
Support vectors
Origin

OCSVM: Kernels

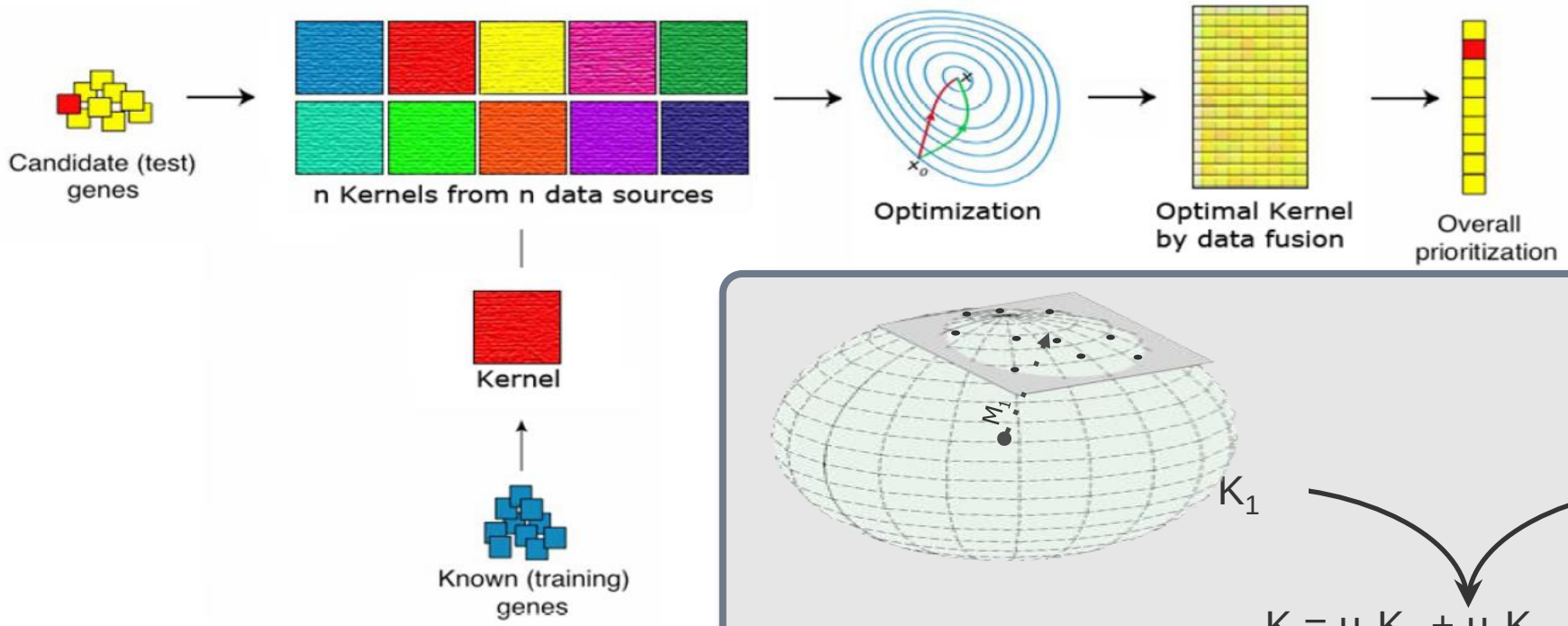
- Equivalent to solving the dual quadratic programming problem
  - $\min_{\alpha} (1/2) \sum_{i,j} \alpha_i \alpha_j K(x_i,x_j)$ s.t. $0 \leq \alpha_i \leq 1/(\nu l)$, $\sum_i \alpha_i = 1$
    - where α_i - Lagrange multiplier

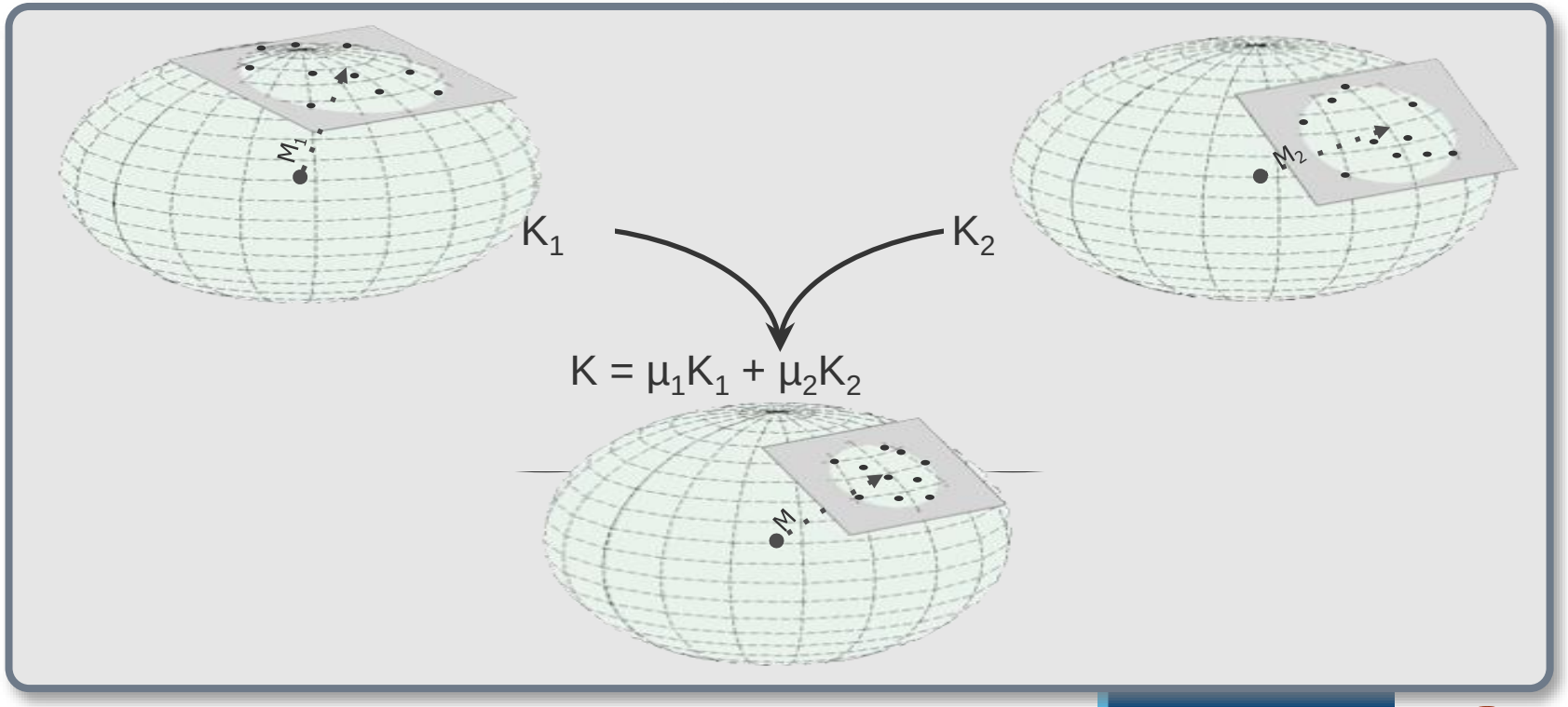
      ν - parameter to control trade-off between distance of hyperplane from the origin and number of points in training dataset

      l - number of points in training dataset

- Kernel function projects input vectors into a feature space allowing for nonlinear decision boundaries
  - Feature map: $\Phi: X \rightarrow R^N$
  - Kernel Function: $K(x_i,x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$

KERNEL-BASED GENOMIC DATA FUSION using MKL





T. .De Bie *et al.*, "Kernel-based data fusion for gene prioritization," Bioinformatics , 2007

GENE PRIORITIZATION USING LOG-EUCLIDEAN MEAN

Incorporating sequence evolution information

PSFM

STRING

Tanimoto

Diffusion

BLAST

UniProt the universal protein resource

GO the Gene Ontology

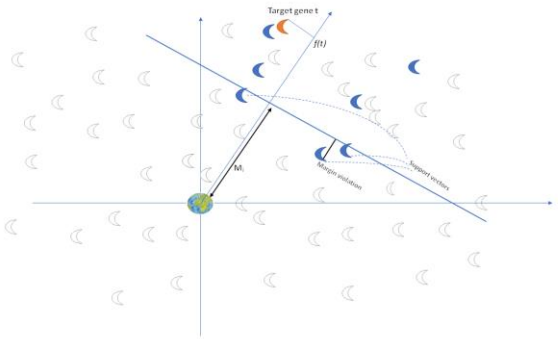
PubMed

β_1 β_2 β_3 β_4 β_5 β_6 β_7

LogE Mean

Combined Kernel

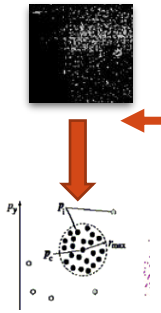
Seeking the hyperplane that best separates the training genes from the origin.



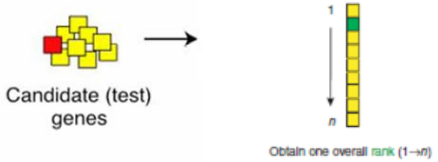
$$K_{LE}(K_1,\ldots,K_n)=\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(K_i)\right)$$

$$K_{LE}(K_1,\ldots,K_n)=\exp\left(\frac{1}{n}\sum_{i=1}^{n}(\beta_i)\log(K_i)\right)$$

$$\sum_{i=1}^{n}\beta_i=1$$



Novelty Detection Using One-class SVM



Ranking Using decision values of One-class SVM Model

Better hunter can access a wider spectrum

KU LEUVEN

GENE PRIORITIZATION: PROSPECTIVE BENCHMARK

- Validate gene prioritization machines in a more realistic task setting.
  - Corss-validation
  - This often leads to an overestimate of the real performance
- A prospective benchmark based on the OMIM associations
  - train the model based on disease-associated genes discovered before a certain time using genomic data sources released prior to that time (2010)
  - (ii) evaluate the model on the prediction of disease-associated genes reported afterwards (between 2010 and 2013)

  - 80 diseases that have at least
    - 3 genes in the 2010 version (727 genes listed in 2010)
    - and have at least one gene reported after 2010 (219 genes reported after 2010)

GENE PRIORITIZATION: PROSPECTIVE BENCHMARK

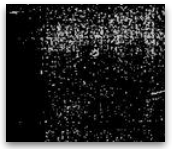
2010

Training

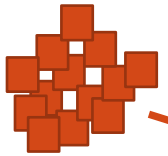
Evaluation

Genomic data sources

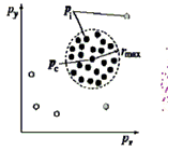
UniProt the universal protein resource | GO the Gene Ontology | BLAST | STRING | PSFM | PubMed



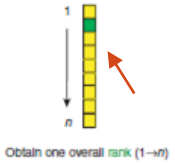
Combined Kernel



Disease- Known Genes

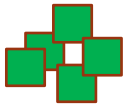


[A2M, AD9, SORL1, HFE, ACE, BLMH, APP, PACIP1, PSEN1, APOE, PSEN2, MPO, NOS3, APBB2, PLAU]



Obesity

disease-associated genes reported afterwards



FFAR4

GENE PRIORITIZATION: RESULTS

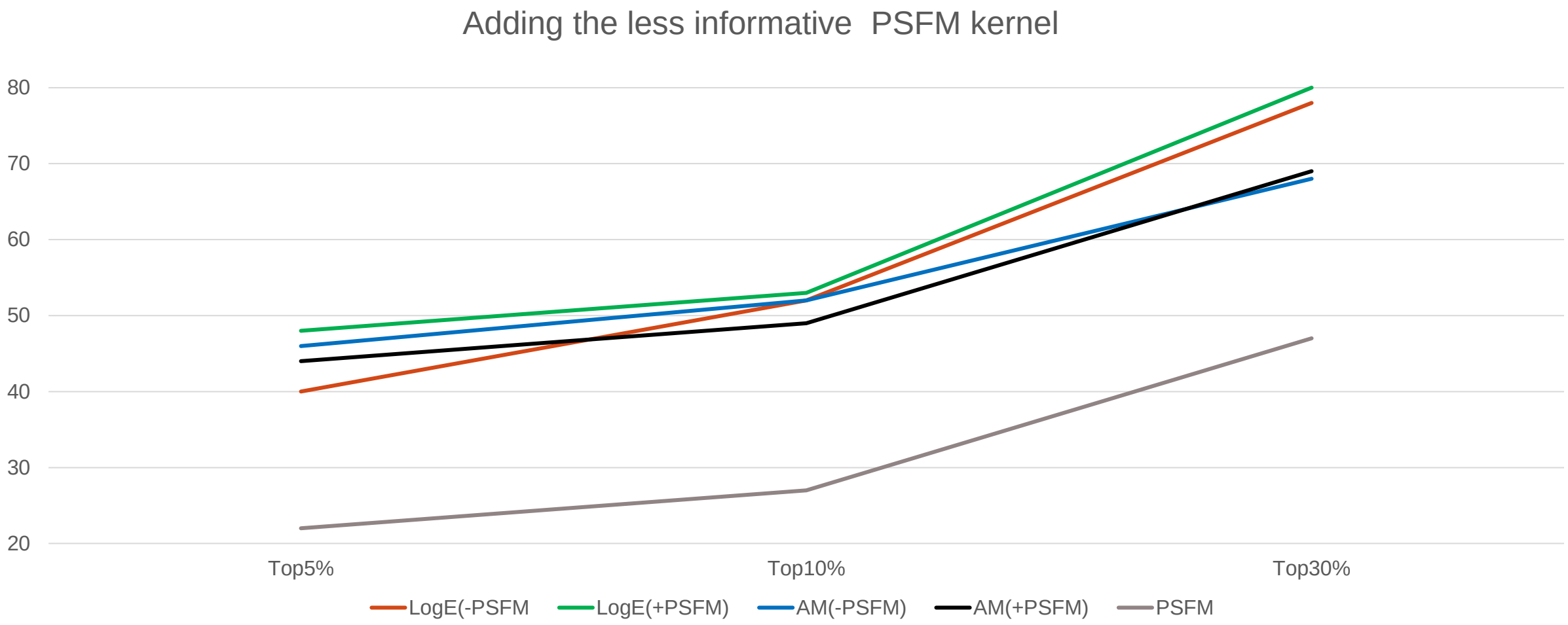
Methods	TPR results for at top of prioritized genes		
	TPR in top 5%	TPR in top 10%	TPR in top 30%
LogE	40%	52%	78%
W-LogE	47%	54%	79%
AM	46%	52%	68%
Endeavour	40%	49%	71%
PSFM	22%	27%	44%

Methods	TPR results for at top of prioritized genes after adding PSFM kernel			p-value of gene ranking
	TPR in top 5%	TPR in top 10%	TPR in top 30%	
LogE	42%	53%	80%	2.1813e-18
W-LogE	49%	55%	80%	5.9909e-20
AM	44%	49%	69%	9.0616e-15

- The linear-based averaging of kernel matrices leads to mixed results if we add a less informative kernel.
- PSFM kernel still carries complementary information when it is integrated with other biological data in an intelligent way, which could improve the performance of the gene prioritization task.

P. Zakeri, et al., "Gene prioritization through geometric-inspired kernel data fusion", IEEE IBIBM, 2015;

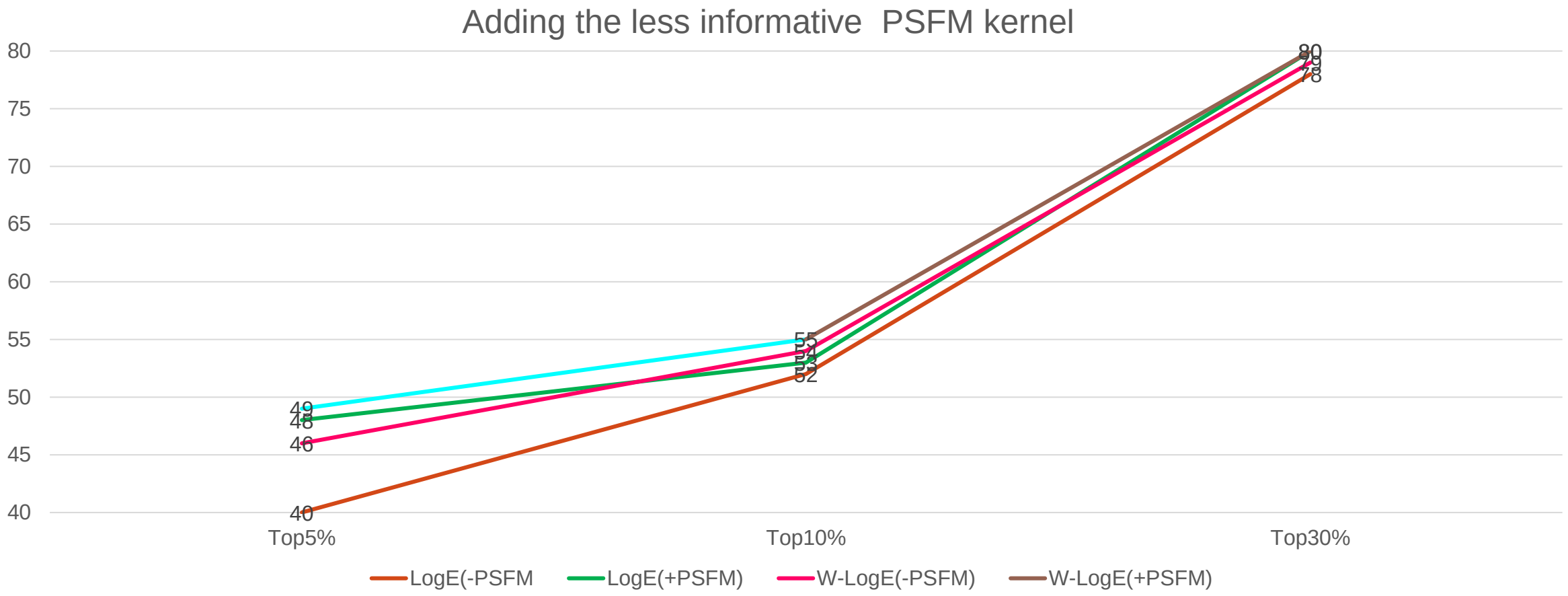
THE LINEAR-BASED AVERAGING OF KERNEL MATRICES LEADS TO MIXED RESULTS IF WE ADD A LESS INFORMATIVE KERNEL



P. Zakeri, et al., "Gene prioritization through geometric-inspired kernel data fusion", IEEE IBIBM, 2015;

KU LEUVEN

LOG-E VS WLOG_E WHEN INCORPORATING A LESS INFORMATIVE KERNEL



P. Zakeri, et al., "Gene prioritization through geometric-inspired kernel data fusion", IEEE IBIBM, 2015;

KU LEUVEN

OUTLINES

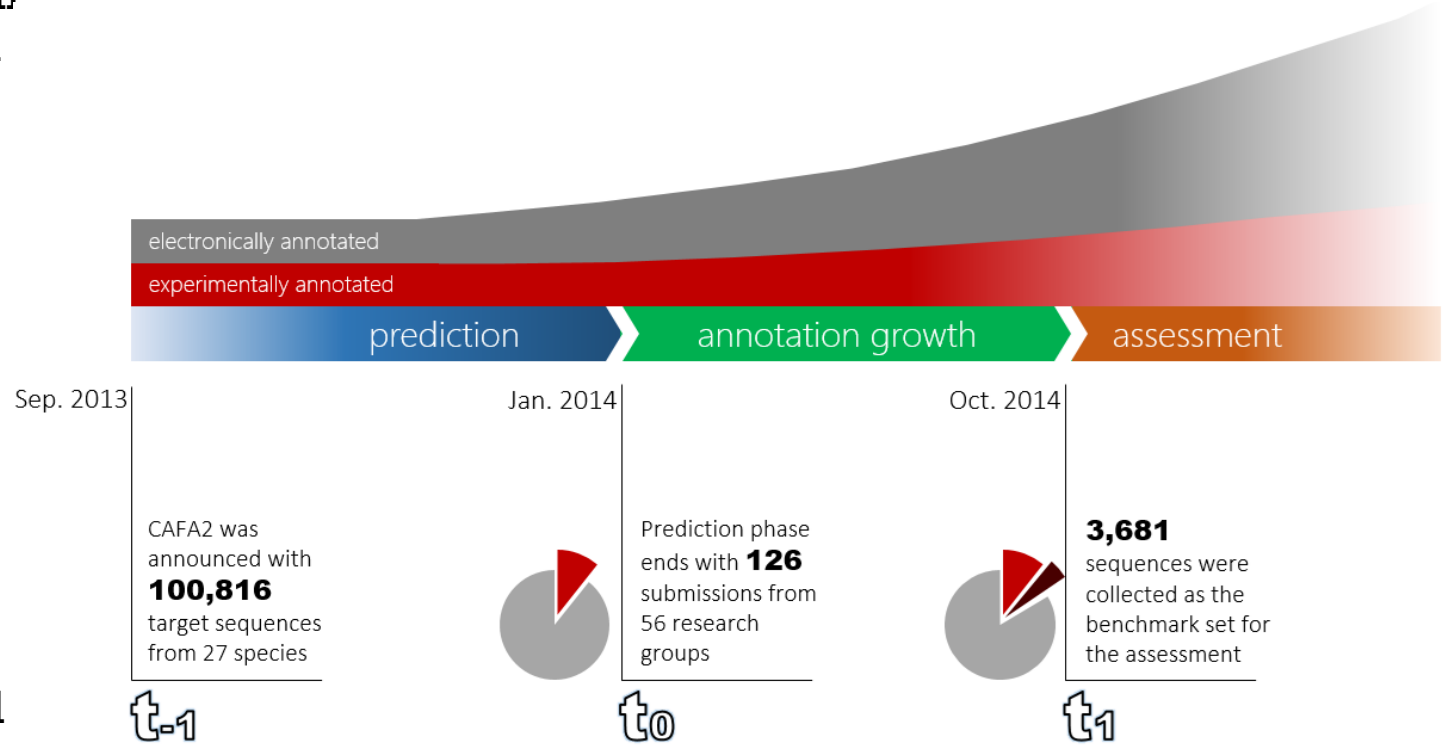
- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

KU LEUVEN

CAFA2 CHALLENGE: AUTOMATED PROTEIN FUNCTION PREDICTION

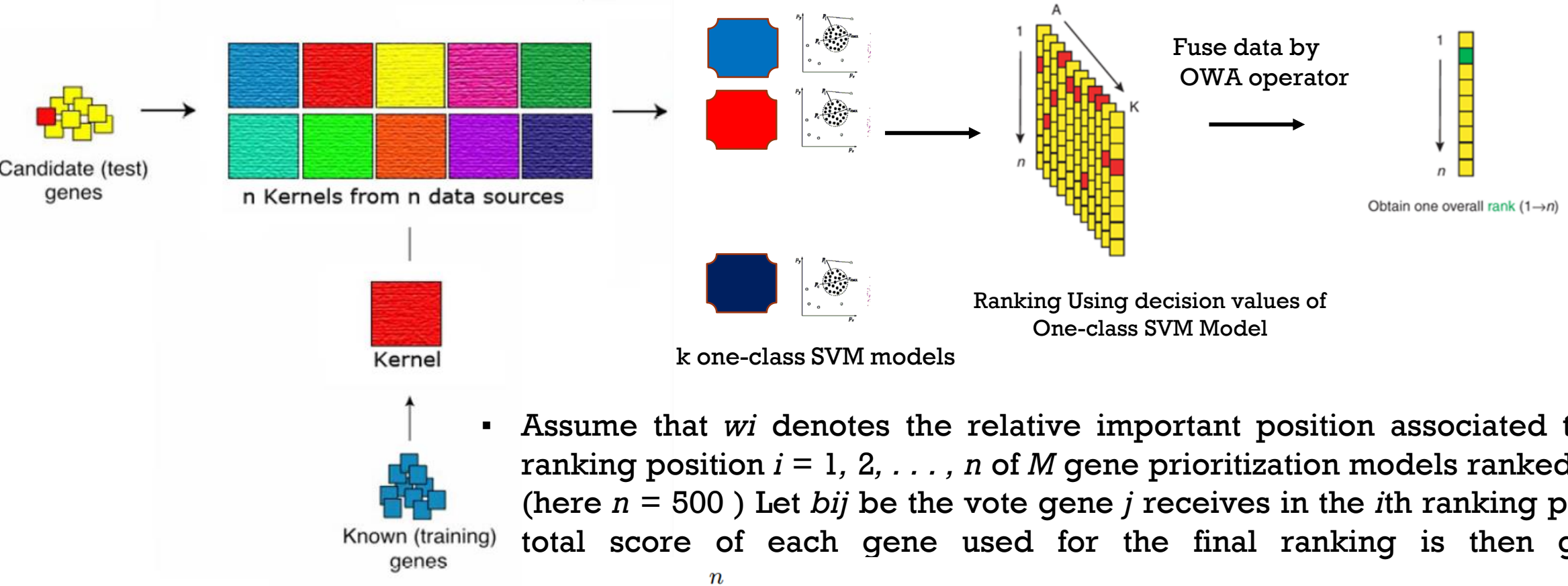
- Second Critical Assessment of Functional Annotation (CAFA)
  - A timed challenge to assess computational methods that automatically assign protein function.
  - Predicting the Molecular Function
  - Predicting the Biological Process
  - Predicting the Cellular Compartment
  - Predicting the Human Phenotype
  - 126 methods from 56 research groups

  - Our model: biological data fusion at the decision level
  - Modeling multi-heterogeneous biological data fusion in the prioritization task based on the ordered weighting averaging



KU LEUVEN

KERNEL-BASED GENOMIC DATA FUSION



- Assume that *wi* denotes the relative important position associated to the *i*th ranking position *i* = 1, 2, . . . , *n* of *M* gene prioritization models ranked *N* genes (here *n* = 500) Let *bij* be the vote gene *j* receives in the *i*th ranking place. The total score of each gene used for the final ranking is then given by

$$R(g_j) = \sum_{i=1}^{n} w_i b_{ij}, \quad j = 1, 2 \ldots, m. \qquad w_n = \frac{6(n-1}{n(n+1)}\left(\frac{2n-1}{3(n-1)} - \beta\right).$$

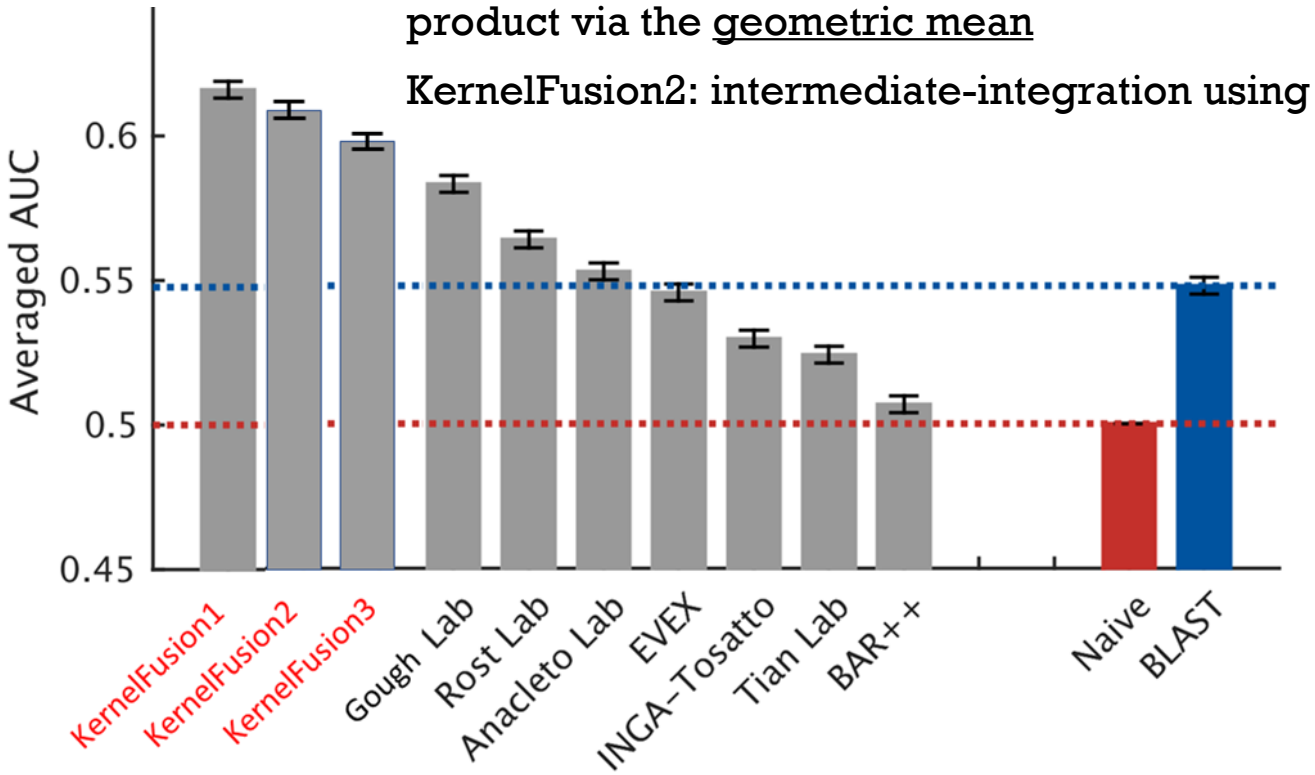
OVERALL EVALUATION USING THE AVERAGED AUC OVER TERMS WITH NO LESS THAN 10 POSITIVE ANNOTATIONS

- Using our first proposed model, about 92% of all HPO terms (494 out of 537 terms) achieve AUCs greater than the averaged AUC over all submitted methods.
- In predicting 103 HPO terms, our first KernelFusion method achieves the highest average AUCs among all participating methods.
- Our second and third KernelFusion methods succeed in obtaining the greatest average AUCs for 70 and 75 HPO terms respectively.

KernelFusion1: late-integration using OWA

KernelFusion3: late-integration using rank product via the geometric mean

KernelFusion2: intermediate-integration using AM



Jiang Y, *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy", Genome Biol., 2016

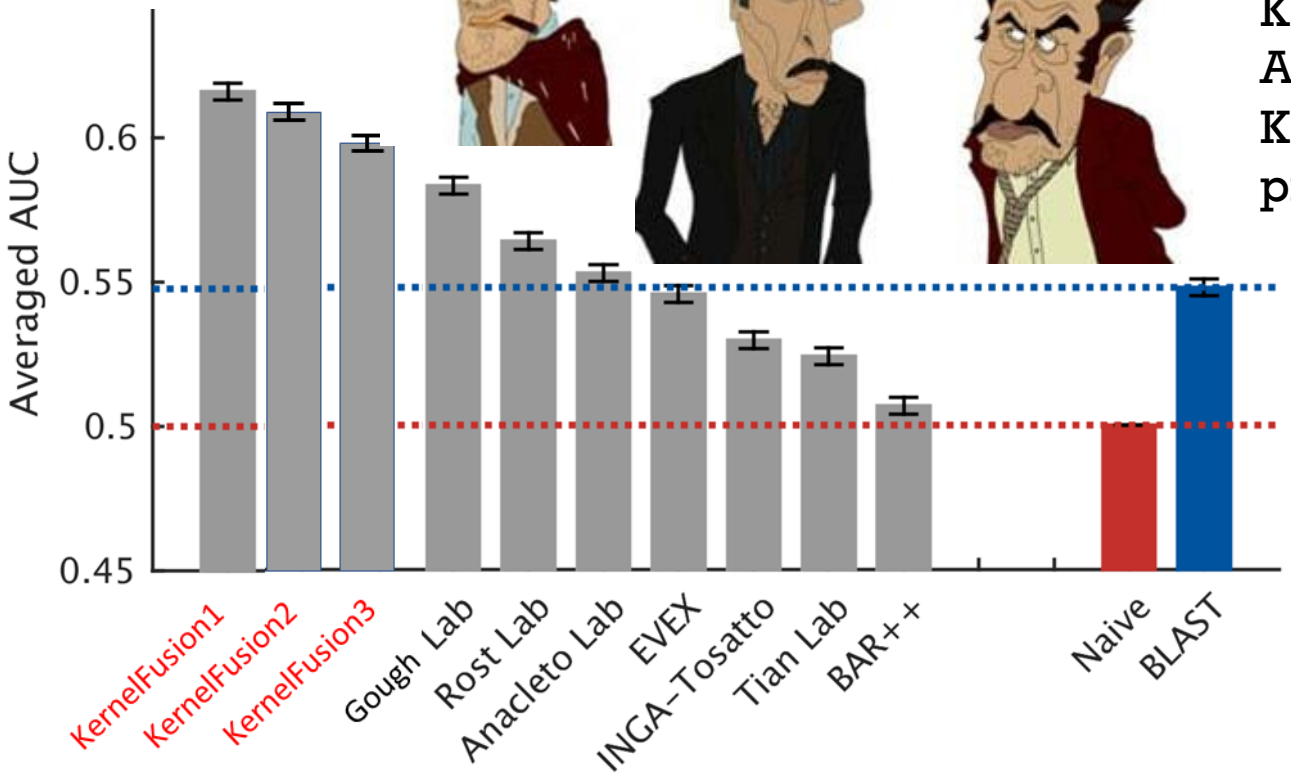
KU LEUVEN

OUTLINES

- A real challenge in bioinformatics
- Concepts and methods for genomic data fusion
  - Investigating the advantage of genomic data fusion at different levels of data realization using Kernel methods
  - Recap SVM
  - Multiple kernel learning
- Averaging IS beautiful, but what do we mean by "average"? Each type of data object needs a proper version of that.
- Geometric kernel data fusion
  - Tackling the protein fold recognition problem for 27 folds
  - Scalable methods for geometric kernel data fusion
  - Combining the evolutionary and secondary structural information could be crucial to elucidate the relationship between primary and tertiary structure in proteins
  - Application of GKF in gene prioritization
  - One-class SVM
- Integrating kernels at decision level
  - Application of late kernel integration in annotating the HPO terms for human protein sequence is a difficult task
    - CAFA Challenge: The best of the worst

OVERALL EVALUATION USING THE AVERAGED AUC OVER TERMS WITH NO LESS THAN 10 POSITIVE ANNOTATIONS

The best of the worst



KernelFusion1: late-integration using OWA

KernelFusion2: intermediate-integration using AM

KernelFusion3: late-integration using rank product via the geometric mean

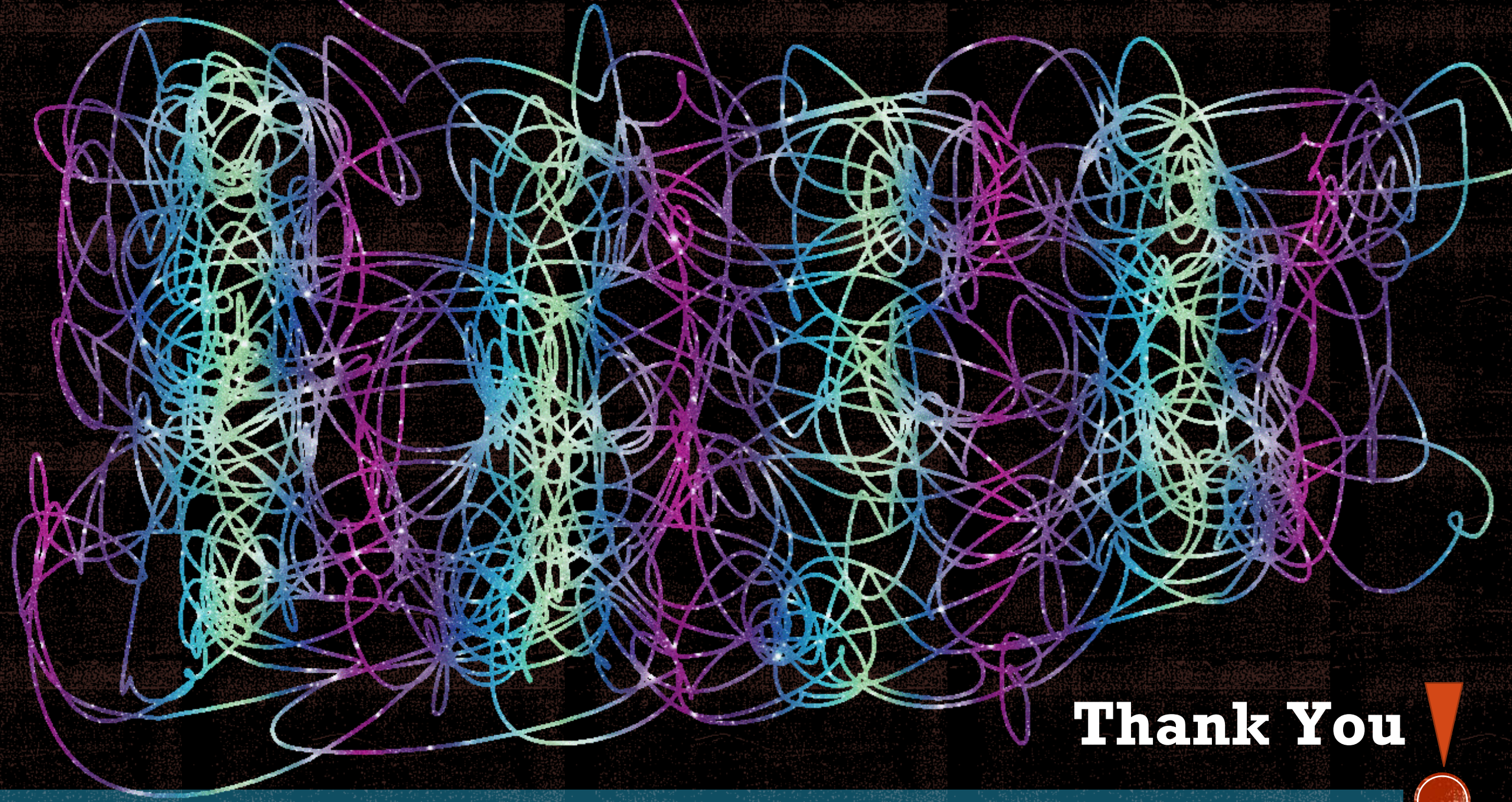
Jiang Y, *et al.*, "An expanded evaluation of protein function prediction methods shows an improvement in accuracy", Genome Biol., 2016

KU LEUVEN

Summary: PROS AND CONS OF KERNEL DATA FUSION METHODS

Kernel fusion strategies	Space com-plexity	Time com-plexity	Learning com-plexity	Handling hetero-geneity	Model and biological interpreta-tiveness	Dependency on learning structure
Raw Fusion	High	Fast	Easy	Does not support	Passively interpretive	Fully dependent
Transitional Fusion	Moderate	Moderate[13]	Moderate	Support	Actively interpretive	Partially dependent
Decision Fusion	Low	Moderate[14]	Moderate	Support	Moderately interpretive	Independent

13) It could be slow in the case of some MKL methods.
14) It could be slow in the case of employing a complex aggregation methods or increasing the number of models participated in decision making.



Thank You !