Time-Series Alignment by Non-negative Multiple Generalized Canonical Correlation Analysis

Bernd Fischer, Volker Roth, and Joachim M. Buhmann

Abhishek Tripathi
Department of Computer Science
University of Helsinki

Quick Overview

- Liquid chromatography coupled to mass spectrometry is widely used for quantitative protein analysis
- A LC/MS device generates mass peaks along time axis
- Non-linear time deformation is a major problem when comparing two biological samples or repeated experiments
- A technique based on Generalized Canonical Correlation Analysis is proposed to align the time series

Motivation

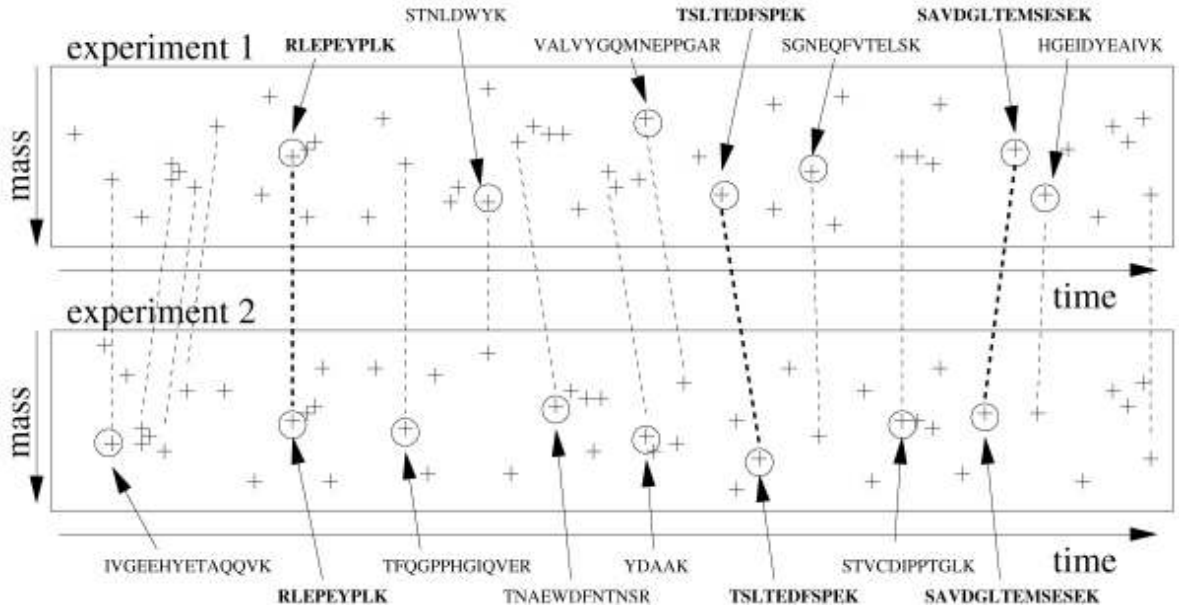
- In quantitative proteomics, it is of particular interest to
  - Classify a protien sample according to some phenotype, e.g. Cancer or non-cancer?
  - Identify relevant proteins discriminating different biological conditions
- Differential protein expression is the answer
  - Proteins are digested into peptides
  - Differential protein expression is estimated over all peptides that correspond to a particular protein
- Absolute expression level can not be robustly measured
  - Unknown ionization efficiency and digestion rates
  - Only differential protein expression can be reliable estimated
- *Problem: Reliable correspondence between peptide measurements in several replicated samples?*

Liquid Chromatography/Mass Spectrometry

- Peptides' amount as a list of peaks in 2D image
  - Mass/charge
  - Time co-ordinate
- Time corresponds to the retention time: when peptide ion elutes from LC columns
- Similar peptides elute within small time window
- Mass axis usually well conserved, but the time axis shows non-linear deformations
- For some peaks, the underlying peptide sequence is known

LC/MS cont.....

- For each experiment, we have at various time points
  - A large list without knowledge of underlying peptide sequence(2000-3000 peaks)
  - A moderate list with known peptide sequence (100-200 peaks)
- The overlap between known peaks between experiments is small
- *The idea is to increase the number of identified peaks by aligning all replicates of the experiments*



Standard Methods for Alignment

- Correlation optimized warping : piece-wise linear functions to align pairs of time series
- A hidden Markov model by *Listgarten et al. 2005*
- Hierarchical clustering for alignment by *Tibshirani et al. 2004*
- Robust point matching by *Kirchner et al. 2007*
- Semi-supervised nonlinear ridge regression by *Fischer et al. 2006*
- Current work extends the idea of *Fischer et al. 2006* by using Generalized Canonical Correlation Analysis
  - Non symmetricity of ridge regression
  - Aligning multiple time series instead of only a pair of time series

Formal Problem Description

- Align K different time scale, each time is a list of peaks with time coordinates

$$P_k = \left\{ t_1^{(k)}, \ldots, t_{n_k}^{(k)} \right\}$$

- A set of known correspondence points between time scales k and l, peptides that are identified in both samples

$$C_{k,l} = \left\{ \left(t_1^{(k)}, t_1^{(l)} \right), \ldots, \left(t_m^{(k)}, t_m^{(l)} \right) \right\}$$

- Determine a mapping $f_{k,l} : P_k \to P_l \bigcup \{\phi\}$, i.e., for a peak $p \in P_k$ find a corresponding peak (if exists) $q \in P_l$
- ϕ represents the case when no corresponding peak is found
- Find a continuous transformtion $g_{k,l} : \Re \to \Re$, transforming the time scale k into the time scale l
- Given the transformation $g_{k,l}$, we create a mapping $f_{k,l}$

$$f_{k,l}\left(t_j^{(k)}\right) = \begin{cases} \arg\min_{t_i^{(l)} \in P_l} \{d_{ij}\} & \text{if } \exists i : d_{ij} \le w \quad \text{where } d_{ij} = \left| t_i^{(l)} - g_{k,l}\left(t_j^{(k)}\right) \right| \\ \varnothing & \text{else.} \end{cases}$$

Estimating Time Transformation Function $g_{k,l}$

- Robust Ridge Regression
  - Let $(x_i, y_i) = (t_i^{(k)}, t_i^{(l)}) \in C_{k,l}$ time correspondence between time series k & l
  - Transform time to polynomial basis $\phi(x_i) = (1, x_i, x_i^2, ..., x_i^d)^t$
  - $\phi(x_i)$: zero mean and unit variance
  - Find parameter vector β that minimizes

$$\sum_{i=1}^{n} L_c(\phi(x_i)^t \beta - y_i) + \lambda \beta^t \beta$$

- Disadvantages of Robust Ridge Regression
  - Unsysmmetric, $g_{k,l}$ is not inverse of $g_{l,k}$
  - Non monotonicity of time transformation function

- *Canonical Correlation Analysis solves these issues*

Canonical Correlation Analysis

- A method of correlating linear relationships between two multidimensional variables
- Let $x = (x_{1,} x_2, ..., x_n)$ and $y = (y_{1,} y_2, ..., y_n)$
- Find the directions W_x and W_y such that

$$\rho = \max_{W_x, W_y} corr(P_x(W_x), P_y(W_y))$$

- Where $P_x(W_x)$ and $P_y(W_y)$ are projections of x and y onto W_x and W_y

$$P_x(W_x) = (\langle W_x, x_1 \rangle, \langle W_x, x_2 \rangle, ..., \langle W_x, x_n \rangle) \text{ and } P_y(W_y) = (\langle W_y, y_1 \rangle, \langle W_y, y_2 \rangle, ..., \langle W_y, y_n \rangle)$$

Computing $g_{k,l}$ using CCA

- Find β_1 and β_2 such that $\max_{\beta_1,\beta_2} corr(\phi(x_i)^t\beta_1, \phi(y_i)^t\beta_2)$
- or $\text{maximize} \frac{\sum_{i=1}^n \beta_1^t \phi(x_i)\phi(y_i)^t \beta_2}{\sqrt{\sum_{i=1}^n (\phi(x_i)^t\beta_1)^2 \sum_{i=1}^n (\phi(y_i)^t\beta_2)^2}}$
- or $\text{minimize} \sum_{i=1}^n (\phi(x_i)^t\beta_1 - \phi(y_i)^t\beta_2)^2 \quad \text{s.t.} \ \|\beta_1\| = 1, \|\beta_2\| = 1.$
- Now, we have $g_k(x_i) = \phi(x_i)^t\beta_k$
- Non-negativity(monotonically increasing time transformation) not yet achieved!

Monotonically increasing Time Tranformation

- Use a set of hyperbolic tangent basis functions

$$\phi(x_i) = \begin{pmatrix} \tanh(\sigma(x_i - z_1)) \\ \tanh(\sigma(x_i - z_2)) \\ \vdots \\ \tanh(\sigma(x_i - z_d)) \end{pmatrix}.$$

- Non-negativity constraint on the regression parameters $\beta_{k,j} \geq 0$
- The cost function now,

$$\text{minimize} \sum_{i=1}^{n} (\phi(x_i)^t \beta_1 - \phi(y_i)^t \beta_2)^2 \quad \text{s.t. } \|\beta_1\| = 1, \|\beta_2\| = 1, \beta_{k,j} \geq 0.$$

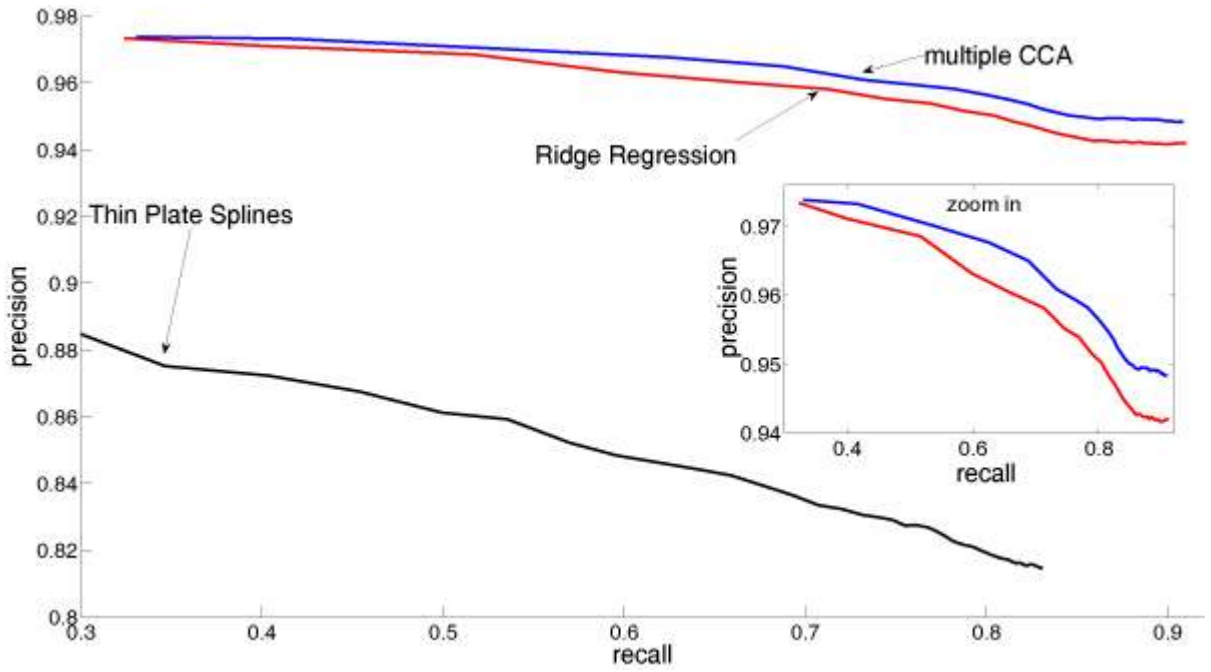
- Solved iteratively by gradient descent

Results

- Data
  - 3 different samples A, B, C from *Araidopsis Thaliana*
  - Sample pair with samples consisting pool(A/B) and pool(B/C)
  - 3 technical replicates of each sample
  - Multiple CCA is used to joinlty align all 6 experiments
  - Results are compared with
    - Robust ridge regression for (6x5)/2 possible pairs
    - Method based on Thin plates spline
- Validation of peak matching with known peptide sequence
- Validation of differential protein expression values

Validation of peak matching with known peptide sequence

- 10 fold cross validation, 9:1 training to test set
- Recall $$(\text{recall})\ rec = \frac{\#correct + \#wrong}{\#correct + \#wrong + \#nomatch}$$
- Precision $$(\text{precision})\ prec = \frac{\#correct}{\#correct + \#wrong}$$



Validation of differential protein expression values

- Technically different samples, no biologically different samples available
- Compute mean log peptide abundance ratio averaged over all peptides for a particular protein
- Protein over/under –expressed between two conditions if average log ratio deviates with t-test significance level α from zero