# TABLE OF CONTENTS

**Volume 1  Number 1  November 2009**

**MicroPath—A pathway-based pipeline for the comparison of multiple gene expression profiles to identify common biological signatures**  
M. Khan, C. B. Gorle, P. Wang, X. H. Liu, S. L. Li………………………………………………………………………………1

**Assessment of bone condition by acoustic emission technique: A review**  
S. Shrivastava, R. Prakash………………………………………………………………………………………12

**Descriptively probabilistic relationship between mutated primary structure of von Hippel-Lindau protein and its clinical outcome**  
S. M. Yan, G. Wu…………………………………………………………………………………………………23

**New blind estimation method of evoked potentials based on minimum dispersion criterion and fractional lower order statistics**  
D. F. Zha…………………………………………………………………………………………………………33

**Visualization of protein structure relationships using constrained twin kernel embedding**  
Y. Guo, J. B. Gao, P. W. H. Kwan1, K. X. S. Hou………………………………………………………………………40
Biosciences

Journal Information

SUBSCRIPTIONS


E-mail: service@scirp.org

**Subscription rates:** Volume 2 2009
Print: $50 per copy.
To subscribe, please contact Journals Subscriptions Department, E-mail: service@scirp.org

**Sample copies:** If you are interested in subscribing, you may obtain a free sample copy by contacting Scientific Research Publishing, Inc at the above address.

SERVICES

**Advertisements**
Advertisement Sales Department, E-mail: service@scirp.org

**Reprints (minimum quantity 100 copies)**
E-mail: service@scirp.org

COPYRIGHT

Copyright © 2009 Scientific Research Publishing, Inc.

All Rights Reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as described below, without the permission in writing of the Publisher.

Copying of articles is not permitted except for personal and internal use, to the extent permitted by national copyright law, or under the terms of a license issued by the national Reproduction Rights Organization.

Requests for permission for other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works or for resale, and other enquiries should be addressed to the Publisher.

Statements and opinions expressed in the articles and communications are those of the individual contributors and not the statements and opinion of Scientific Research Publishing, Inc. We assume no responsibility or liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained herein. We expressly disclaim any implied warranties of merchantability or fitness for a particular purpose. If expert assistance is required, the services of a competent professional person should be sought.

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: biosci@scirp.org
MicroPath-A pathway-based pipeline for the comparison of multiple gene expression profiles to identify common biological signatures

Mohsin Khan, Chandrasekhar Babu Gorle, Ping Wang, Xiao-Hui Liu, Su-Ling Li

ABSTRACT
High throughput gene expression analysis is swiftly becoming the focal point for deciphering molecular mechanisms underlying various different biological questions. Testament to this is the fact that vast volumes of expression profiles are being generated rapidly by scientists worldwide and subsequently stored in publicly available data repositories such as ArrayExpress and the Gene Expression Omnibus (GEO). Such wealth of biological data has motivated biologists to compare expression profiles generated from biologically-related microarray experiments in order to unravel biological mechanisms underlying various states of diseases. However, without the availability of appropriate software and tools, they are compelled to use manual or labour-intensive methods of comparisons. A scrutiny of current literature makes it apparent that there is a soaring need for such bioinformatics tools that cater for the multiple analyses of expression profiles.

In order to contribute towards this need, we have developed an efficient software pipeline for the analysis of multiple gene expression datasets, called MicroPath, which implements three principal functions; 1) it searches for common genes amongst n number of datasets using a number crunching method of comparison as well as applying the principle of permutations and combinations in the form of a search strategy, 2) it extracts gene expression patterns both graphically and statistically, and 3) it streams co-expressed genes to all molecular pathways belonging to KEGG in a live fashion. We subjected MicroPath to several expression datasets generated from our tolerance-related in-house microarray experiments as well as published data and identified a set of 31 candidate genes that were found to be co-expressed across all interesting datasets. Pathway analysis revealed their putative roles in regulating immune tolerance. MicroPath is freely available to download from: www.1066technologies.co.uk/micropath.

Keywords: Co-Expression Analysis, Microarray, Permutations and Combinations, Multiple Gene Expression Analysis

1. INTRODUCTION
There is a general consensus amongst scientists and researchers that the fundamental asset of microarray technology lies in its inherent ability to produce a global snapshot of the cellular state in the milieu of any given biological question. It is therefore not surprising that microarrays have revolutionised the field of molecular biology by offering an efficient and cost effective medium for biologists to quantify mRNA transcript levels of several thousands of genes concurrently in order to observe specific states of the transcriptome (in response to a particular treatment or specific time point). Owing to this innate faculty to decipher the transcriptome, gene expression profiles pertaining to a wide variety of biological questions are being rapidly generated by scientists worldwide and are deposited and subsequently made accessible through public repositories such as ArrayExpress [1] and the Gene Expression Omnibus [2]. With so much wealth of high throughput biological data made available, biologists have become motivated to utilise these sets of data in an attempt to investigate common regulatory signatures, which may be implicating the transcriptome state across multiple gene expression profiles sharing a similar biological theme. One of the most widely accepted methodologies of comparing expression profiles is based on the assumption that genes across different biological conditions sharing similar expression patterns are likely to be involved in the same biological processes [2], and therefore, may share common regulatory signatures. By using this method of comparison, which is one of the most successful methods to date, coupled with the availability of publicly available data
repositories offering gene expression profiles, biologists have been granted the opportunity to answer complex biological questions pertinent to biological phenomena underlying various different disease states.

To this end, we have developed a novel bioinformatics software pipeline called MicroPath, which specialises in the cross comparison of multiple gene expression datasets and attempts to identify common regulatory signatures from the standpoint of molecular pathway analysis. When one scrutinises current literature relevant to automated solutions of gene expression analysis, it becomes apparent that there is an increasing demand for software applications that offer an efficient pipeline to the analysis of multiple gene expression profiles. Although current meta-analyses studies have been conducted with the purpose of employing statistical techniques to compare cDNA and affymetrix gene expression profiles [3,4,5,6], it cannot be denied that there is a mounting need for this process to be automated. Nevertheless, various approaches/algorithms of statistical nature have already been implemented with the purpose of identifying the most relevant pathways in a given experiment [7,8,9] together with methods such as gene set enrichment analysis (GSEA), which ranks genes based on the correlations between their expressions and observed phenotypes in the context of biological pathway discoveries [10]. There are also tools available that functionally annotate gene expression data [11,12]. Albeit, it remains infeasible for biologists to cross compare several expression profiles without an automated solution, and hence, they are faced with the labour-intensive task of employing manual methods to carry out comparisons. MicroPath uses the meta-analytic standard and has been specifically developed to: compare several significantly expressed sets of genes in order to find the intersection of common genes using both number crunching methods as well as the classical permutation and combination principle, extract putative regulatory signatures using both statistical and graph-based approaches and finally, mapping these sub-sets of co-expressed genes to molecular pathways all in the form of a high throughput pipeline.

2. IMPLEMENTATION

The front-end of MicroPath was developed in Visual Basic.Net and Perl, and the database back-end was developed in MySQL. Upon analysing the users input files (gene expression profiles), processed data is displayed intuitively on the graphical user interface, which is equipped with various interactive objects such as charting facilities, buttons, drop-down menus and user input/output dialogues. The interface is also equipped with a function to export processed data into Microsoft excel for further scrutiny and use.

2.1. System Architecture

MicroPath carries out meta-profiling of multiple gene expression datasets using two different approaches. Firstly, the intersection of common genes is identified across n number of expression profiles, which is then plotted graphically using a simple number crunching exercise. The second approach applies to a situation where an attempt to identify common genes across n number of expression profiles using the aforementioned approach fails due to the absence of common genes across all datasets (this situation is especially common when a large number of expression profiles are compared, which reduces the probability of finding a common gene amongst them). Consequently, MicroPath applies the permutations and combinations mathematical principle to solve this problem (refer to implementation of meta-analysis strategy below for details). Once the intersection of a set of common genes has been identified and subsequently displayed on the interface (using either of the above methods), the next stage in the analysis is to extract patterns from the intersection in order to identify common genes that are being expressed in accordance with the biological question. MicroPath offers a semi-automated graph-based approach to achieve this as well as classical statistics to identify the overall correlation of gene expression. Finally, co-expressed genes (common genes that are expressed in accordance to the relevant biological question) are mapped to all molecular pathways known to date in order to reveal their molecular dependencies (refer to Figure 1 for the complete system architecture).

2.2. Implementation of Meta-analysis Strategy

In theory, an intersection of a sub-set of common genes across multiple gene expression profiles should be easily attainable using simple number crunching methods of comparison. In practice, this is not always the case since the likelihood of identifying genes sharing common accession identifiers decreases as the number of profiles to compare increases. This inverse relationship makes sense both mathematically and biologically. From a biological perspective, regulatory signatures tend to be diluted over entire datasets and as a result, only a proportion of the total number of profiles to compare may actually share common genes. In such a scenario, using a simple method of comparison would break down at some point and no common genes would be reported to the user, although common genes may be present within n-1 expression profiles. To prevent potentially interesting biological findings to be hampered at this point in the analysis, we have applied the principle of mathematical combinations to the comparison of multiple gene expression profiles. All possible combinations of comparing n number of datasets with each other are firstly computed using the combination equation:

\[
\binom{n}{r} = \frac{n!}{r!(n-r)!}
\]

Where nf is the factorial of the total number of datasets n, and rf is the factorial of the selected number of datasets to compare when comparing n datasets results in zero common genes, r.
This generates the total number of permutations of comparing datasets (Cr) for given values of n (total number of datasets imported by user) and r (number of intended datasets used to search for common genes when zero common genes are reported across n datasets) (Table 1).

Figure 1. Functions of MicroPath. Users are prompted to import up to 10 gene expression profiles, which are then compared using a direct comparison method. If this method yields zero common genes, MicroPath automatically attempts to identify an intersection of common genes by reducing the search space to n-1 datasets using permutations and combinations. This process is continued until at least 1 common gene is reported. Following this, users are provided with a function to search for expression patterns graphically and gene expression correlations are calculated statistically using the pearson’s correlation coefficient algorithm. Finally, co-expressed genes are mapped to all molecular pathways of KEGG in a high throughput fashion by automatically accessing its API via SOAP-Lite.
Table 1. Multiple gene expression profile search strategy generated from applying the principle of permutations and combinations. The first column represents the total number of expression datasets, \( n \), that users may import (this is the search space). The second column represents, \( r \), the number of expression datasets to compare if zero common genes are reported to be matched across \( n \) datasets. The final column represents the total number of mathematical combinations possible for each given value of \( n \) and \( r \).

<table>
<thead>
<tr>
<th>Total number of expression datasets (( n ))</th>
<th>Number of intended expression datasets to compare when comparing ( n ) datasets yields no results (( r ))</th>
<th>( n - r )</th>
<th>Total number of combinations of ( r ) (( \binom{r}{n} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>3</td>
<td>120</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>4</td>
<td>210</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>5</td>
<td>252</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>6</td>
<td>210</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>7</td>
<td>120</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>8</td>
<td>45</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>6</td>
<td>3</td>
<td>84</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>4</td>
<td>126</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>5</td>
<td>126</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>6</td>
<td>84</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>3</td>
<td>56</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>5</td>
<td>56</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

These combinations of datasets (\( \binom{r}{n} \)) are then used as a criterion to search for common genes across \( r \) number of gene expression profiles when comparing \( n \) number of datasets fail to yield any common genes. However in this scenario, \( n \) number of datasets are still used as the search space from which all possible combinations (\( \binom{r}{n} \)) of \( r \) datasets are compared to each other in order to increase the probability of finding a common gene. Once common genes have been identified using this method, MicroPath will report the results to the interface.

2.3. Extracting Gene Expression Patterns Graphically and Statistically

Following the identification of common genes across \( n \) datasets using either of the methods described earlier, the next stage in the analysis is to generate a graphical representation of this expression data from which biologically meaningful patterns can be extracted. Because signals pertaining to transcriptome states tend to be diluted over entire profiles, a specific criterion is required to narrow down the common genes of interest to include only those genes that are consistently regulated according to the biological question. The assumption we have made is that any given common gene across \( n \) datasets can exhibit one of three specific behaviours. It can either be consistently upregulated across all datasets, downregulated across all datasets and up or downregulated across all datasets. Based on the nature of the specific biological question, users can select the appropriate pattern from the options, which will result in a graphical display of those genes which satisfy the search criteria. Together with this faculty to graphically extract patterns for individual gene expression data points, MicroPath also implements the pearsons correlation coefficient statistical test in order to extract a global gene expression pattern existing between common genes pertaining to two individual expression profiles. The correlations are calculated in a pair-wise manner until each expression data has been statistically compared to all other datasets within \( n \), according to the pearsons correlation coefficient equation:

\[
r = \frac{\sum X Y - (\sum X)(\sum Y)}{n} \sqrt{\frac{\sum X^2 - (\sum X)^2}{n} \frac{\sum Y^2 - (\sum Y)^2}{n}}
\]
Each pair-wise score is then finally averaged in order to provide a global measure of correlation existing between expression profiles. Scores are reported from -1 (perfect negative correlation) to 1 (perfect positive correlation).

2.4. High Throughput Molecular Pathway Analysis

To decipher molecular mechanisms fundamental to the researcher’s biological question, it is necessary to map common gene expression profiles of co-expressed genes to molecular pathways. This is because biological pathways reveal molecular dependencies that exist between genes by illustrating how they collaborate with one another when they participate in specific biological functions. Furthermore, pathways reveal various signalling cascades that play imperative roles in dictating these gene associations. In light of this, we have implemented Micropath to access the Application Programming Interface (API) of the molecular pathway database belonging to KEGG [13] using SOAP-Lite in order to dynamically interact with the static pathway maps. Perl scripts were written for MicroPath to specifically 1) search for user’s co-expressed genes in all biological pathways, 2) highlight genes on to pathways, and 3) return the results of the search to Micropath’s interface (i.e. URL’s of colour coded pathway maps) (Figure 2). Once MicroPath has searched for all of the user’s co-expressed genes in all of the molecular pathways, the URL of each pathway is displayed on the sub-interface. In order to avoid redundancy issues, the URL for each pathway will highlight all co-expressed genes that participate in a given pathway. To help users identify biologically meaningful pathways relevant to their specific biological question, MicroPath will calculate the number of genes identified in a given pathway and 1) express this as a percentage in relation to the total number of common genes from the intersection and 2) express this as a percentage in relation to the total number of genes belonging to that pathway.

Clicking on these links will generate the specific KEGG pathway in HTML on which users co-expressed genes will be highlighted.

Figure 2. Flow diagram of how MicroPath carries out high throughput molecular pathway analysis by connecting to the API of KEGG.
2.5. Generating and Processing Gene Expression Datasets

Gene expression datasets used for the purpose of this article were generated from our in-house microarray experiments as well as published datasets, where the fold change approach was used to select a set of differentially expressed genes from pre-processed data. Matchminer [14] and the Synergizer [15] tools were used to convert gene Hugo identifiers and long names into Genbank accession Id’s in order to ensure that the gene identifiers were of the same type across all datasets prior to comparison. Raw expression data was generated, filtered and normalised using GenePix pro 4.1 [16] and Acuity 4.0 [17] software. Although we used cDNA microarray data for the purpose of demonstrating MicroPath’s capabilities, other data types generated from different platforms such as affymetrix can also be analysed provided Genbank accession identifiers are used to represent the genes.

3. RESULTS AND DISCUSSION

Regardless of the biological question, a typical microarray experiment almost always results in the generation of a set of differentially expressed genes, which represents genes of most importance to the biologist. Therefore, by carrying out several biologically related microarray experiments, several sets of differentially expressed genes would be generated, which would need to be compared and mined efficiently in order to help answer the biological questions asked by the investigators from different research laboratories around the world. Employing manual methods of comparison in this situation would be very inefficient and infeasible. In light of this, to demonstrate the benefits that can be derived from analysing multiple gene expression profiles using MicroPath, we employed datasets generated from our in-house microarray experiments as well as published data. The biological question related to these studies focussed on unravelling the underlying molecular mechanisms dictating immune tolerance by analysing the role of Egr-2 in implicating T-cell tolerance. Although the Early Growth Response gene (Egr-2) has been recently characterised as a candidate tolerance-inducing transcription factor, which interacts with specific genes in order to induce the state of T-cell tolerance [18,19], the possibility of further putative unknown target genes exists that may be vital to the mechanism of tolerance. Hence, the biological purpose of our experiments was to attempt to identify such potentially important genes via the comparison of biologically related expression datasets using MicroPath.

Data consisting of a set of differentially expressed genes generated from the comparison of tolerance Vs activated mice CD4+ T cells was obtained from the ArrayExpress website (accession number: e-mexp-283). The first in-house experiment aimed to generate differentially expressed genes from the comparison of an un-stimulated T cell line from which the Egr-2 gene had been knocked out and a wild type un-stimulated cell line. The second in-house experiment focussed on the comparison between an Egr-2 knock-out T cell line activated with CD3/CD28 for 6 hours and a wild type cell line also activated with CD3/CD28 for 6 hours. Results generated from these experiments were then compared with the aforementioned published tolerance data using MiNer in order to understand the molecular mechanisms controlling immune tolerance.

3.1. Comparison of Gene Expression Profiles Pertaining to Immune Tolerance

The first step in the analysis was to subject the above-mentioned expression profiles to MicroPath in order to identify genes amongst them that had the same accession identifiers. Having done this, MicroPath identified 31 differentially expressed genes that were common to all three expression datasets and generated a graph to delineate their expression values (Table 2, Figure 3). A simple number crunching exercise was used to perform this task since its use generated a reasonable number of common genes, which did not warrant the use of permutations and combinations to perform the search. The next step was to use these 31 differentially expressed genes as a search space to determine those genes that have the potential to be co-expressed. In order to do this, we employed MicroPath’s graphical utility to extract gene expression patterns, which led to the identification of 6/31 genes that were found to be upregulated in tolerance Vs activated CD4+T-cells and downregulated in both p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 datasets (Table 2). The remaining 25 common differentially expressed genes were found to be highly and lowly expressed in tolerance and knock-out datasets respectively. Statistical analysis revealed an overall Pearson’s correlation score of 0.109 from the pair-wise comparison of tolerance data with p-KOA0 Vs WTA0 and a score of -0.123 from the comparison of tolerance with p-KOA6 Vs WTA6. Furthermore, Reverse Transcriptase PCR experiments confirmed that 15 genes from our tolerance Vs activated data were found to be highly expressed in immune tolerance and from these 15 genes, 8 were found to be common amongst all three expression profiles (Table 2).

Because Egr-2 has been previously characterised and found to be highly upregulated in immune tolerance, these results generated from MicroPath are biologically significant because as expected, those genes that were highly expressed in our tolerance Vs activated datasets were found to be insignificantly expressed in our p-KOA6 Vs WTA6 and p-KOA0 Vs WTA0 datasets (from which the Egr-2 gene was knocked out of the cell lines). Amongst these genes, Ap1s1, Slh, Surf6, Vil2, Lilrb4, Tbx21 and Pdcd11lg2 (Table 2) have been confirmed to be upregulated in the process of immune tolerance [20], all of which were found to exhibit low expression values in our knock-out expression datasets. This consistent gene expression pattern can be seen graphically in Figure 3. However, from the 31 interesting common genes, 16 were not confirmed to be involved in
tolerance by RT-PCR yet some of them also exhibited a coherent pattern of gene expression. For example, Ptna, Sed2, Hdac6, Pftp and Chka were all highly expressed in tolerance and conversely downregulated in both knock out datasets. There is a possibility that these genes may also be insignificantly expressed due to the absence of Egr-2. However, conducting RT-PCR for these specific genes would be required in order to confirm that their over-expression results in T-cell tolerance.

3.2. Deciphering Gene Regulatory Networks of Co-Xpressed Genes Via High Throug

The final stage of the analysis entails using MicroPath’s function to connect to the Application Programming Interface (API) of KEGG via SOAP-Lite in order to carry out high throughput molecular pathway analysis. Therefore, for this stage in the analysis, we used MicroPath to map 31 of our co-expressed interesting genes to KEGG pathways and from these 31 genes, 14/31 were identified in a total of 31 molecular pathways (Table 3). Interestingly, several of these pathways were related to the study of immunology and illustrated biological networks such as MapKinase, Jak-Stat, T-cell receptor signalling and Cytokine-cytokine interactions. More specifically, the Pdcd1lg2 gene (accession id: NM_021396) was identified in the Cell Adhesion Molecules (CAM) pathway (Table 3) and studies have confirmed that the over-expression of Pdcd1lg2 has resulted in consistently low levels of Interleukin-2 (IL-2) in naive CD4(+)-T-cells [21]. Further studies have correlated the over-expression of this gene to the negative regulation of T-cell activation. In one particular study, PDL2 (Pdcd1lg2) deficient mice were created in order to characterise the function of this gene in T-cell activation and tolerance, and results generated from this study suggested that Antigen-presenting cells from PDL2-deficient mice were found to be more potent in activating T-cells in vitro when compared to the wild-type counterparts [22]. These findings are conclusive and correlate well with the results generated from our in-house microarray experiments because using MicroPath to compare all three of our datasets followed by extracting gene expression patterns from them resulted in an important finding that Pdcd1lg2 was not only found to be over-expressed in tolerance (fold change of 3.921), but it was also under-expressed in our KOA0 Vs WTA0 and KOA6 Vs WTA6 datasets.

Table 2. Tabulated overview of gene accession ids, Hugo ids and fold change values belonging to 31 common genes identified from the comparison of tolerant Vs activated CD4+ T-cells, p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 expression datasets. Entries highlighted in bold represent genes that were found to be up-regulated in tolerance and conversely downregulated in both p-KOA0 Vs WTA0 and p-KOA6 Vs WTA6 datasets. Entries with * represent genes that have been confirmed to be highly expressed in tolerance by RT-PCR.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>HUGO ID</th>
<th>Fold Change (p-KOA0 Vs WTA0)</th>
<th>Fold Change (p-KOA6 Vs WTA6)</th>
<th>Fold Change (Tolerance Vs Activated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_007381</td>
<td>Acad8 *</td>
<td>0.371336 0.624525</td>
<td>6.373</td>
<td></td>
</tr>
<tr>
<td>NM_007457</td>
<td>Ap1s1 *</td>
<td>0.542474 0.31525 4.965</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_007664</td>
<td>Cd2 4</td>
<td>0.243646 0.7999 1.658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_008305</td>
<td>H2-M9</td>
<td>-0.08048 0.116134 2.857</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_008972</td>
<td>Ptna</td>
<td>-1.31334 -0.46688 5.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_009128</td>
<td>Scd2</td>
<td>-0.18816 -0.39366 4.552</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_009168</td>
<td>Shd *</td>
<td>-0.17495 -0.53582 2.838</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_009298</td>
<td>Surf6 *</td>
<td>0.272072 0.126301 4.365</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_009465</td>
<td>Axl</td>
<td>0.149539 1.475806 3.836</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_009510</td>
<td>Vli2 *</td>
<td>-0.49824 0.319645 3.151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_010102</td>
<td>Edgif</td>
<td>0.311489 0.132689 1.573</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_010413</td>
<td>Hdac6</td>
<td>-0.90335 -0.8226 4.745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_010548</td>
<td>H10 *</td>
<td>3.083863 1.660739 3.521</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_010638</td>
<td>Breh1</td>
<td>0.024803 -0.42543 1.613</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_011325</td>
<td>Pftp</td>
<td>-0.53544 -0.71558 4.363</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_011620</td>
<td>Tnnt2</td>
<td>-0.61646 0.035844 1.665</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_011696</td>
<td>Vdsc3</td>
<td>-0.98084 0.191964 4.701</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_011705</td>
<td>Vtk1</td>
<td>0.460622 -0.34601 2.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_012488</td>
<td>Cd4</td>
<td>0.504494 -0.40277 4.905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_013490</td>
<td>Chka</td>
<td>-2.13728 -0.69458 5.677</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_013532</td>
<td>Lirb4 *</td>
<td>0.79235 1.110898 2.111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_013615</td>
<td>Osr2</td>
<td>2.776384 3.004449 4.809</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_013814</td>
<td>Gait1</td>
<td>-0.47752 0.500297 2.246</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_013866</td>
<td>Zfp385</td>
<td>0.118995 0.428591 1.664</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_016772</td>
<td>Ech1</td>
<td>-0.0666 0.053081 4.284</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_019507</td>
<td>Tbx21 *</td>
<td>0.124767 -0.32731 1.595</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_019561</td>
<td>Ensa</td>
<td>0.778767 -0.44703 1.718</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_019777</td>
<td>Ikbe</td>
<td>0.291602 -0.00772 1.609</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_020027</td>
<td>Bat2</td>
<td>0.291219 -0.23966 5.091</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_021396</td>
<td>Pdcd1lg2 *</td>
<td>1.140087 0.079182 3.921</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NM_021538</td>
<td>Cope</td>
<td>0.154049 0.264541 2.035</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
particular finding is in agreement with the aforementioned studies, concluding that Pdcd1lg2 has a negative inhibitory role towards the process of T-cell activation. In addition, molecular pathway analysis of the Interleukin-10 (IL-10) gene using MicroPath, identified its role in the Cytokine-cytokine interaction, Jak-STAT and T-cell receptor signalling pathways; all three of which are important immunological pathways. IL-10 is a well known cytokine, which has previously been shown to successfully induce immune tolerance in Dendritic Cells [23]. Results generated from MicroPath revealed that IL-10 was highly expressed in our tolerance data with a fold change of 3.521, which was found to be expressed lower in our KOA0 Vs WTA0 profile (fold change: 3.084). Interestingly, following activated with CD3/CD28 for 6 hours, its expression dropped significantly to 1.66, perhaps attributable to the absence of Egr-2. Likewise, other genes from the 31 co-expressed interesting genes show similar patterns of expression and perhaps may be candidate genes for Egr-2 mediated T-cell tolerance. However, this is yet to be confirmed by publications. Finally, the pathway analysis function of MicroPath was used to calculate the percentage of genes identified in each pathway in relation to 1) the intersection of common genes and 2) the total number of genes comprising each pathway. From the results, the Cell Adhesion Molecules (CAM) pathway was particularly significant since 12.91% of the overall pathway was affected by 6.84% of genes common to all 3 expression profiles (Table 4).

![Figure 3. A preliminary graphical overview of common interesting genes generated from the comparison of tolerant Vs activated CD4+ T cells (green), p-KOA0 Vs WTA0 (red) and p-KOA6 Vs WTA6 (blue) expression datasets. It can be seen that genes that are highly expressed in tolerance appear to be expressed poorly in the knock-out datasets. This pattern is consistent throughout the 31 gene expression data points.](image link)

Table 3. Tabulated data generated from high throughput molecular pathway analysis of co-regulated genes. 14/31 common interesting genes were identified in a total of 31 molecular pathway maps of KEGG.

<table>
<thead>
<tr>
<th>GenBank Accession ID</th>
<th>HUGO ID</th>
<th>Pathway ID</th>
<th>Total No of pathways</th>
<th>GenBank Accession ID</th>
<th>HUGO ID</th>
<th>Pathway ID</th>
<th>Total No of pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_007381</td>
<td>Acadl</td>
<td>mmu00071</td>
<td>5</td>
<td>NM_009510</td>
<td>Vll2</td>
<td>mmu04670</td>
<td>2</td>
</tr>
<tr>
<td>NM_007664</td>
<td>Cdh2</td>
<td>mmu04514</td>
<td>1</td>
<td>NM_008205</td>
<td>H2-M9</td>
<td>mmu04514</td>
<td>3</td>
</tr>
<tr>
<td>NM_013488</td>
<td>Cd4</td>
<td>mmu04514</td>
<td>4</td>
<td>NM_013814</td>
<td>Galnt1</td>
<td>mnu00512</td>
<td>2</td>
</tr>
<tr>
<td>NM_011696</td>
<td>Vdac3</td>
<td>mnu04020</td>
<td>1</td>
<td>NM_019777</td>
<td>Ikbke</td>
<td>mnu04010</td>
<td>2</td>
</tr>
<tr>
<td>NM_011125</td>
<td>Ptp</td>
<td>mnu03320</td>
<td>1</td>
<td>NM_010102</td>
<td>Edg6</td>
<td>mnu04080</td>
<td>1</td>
</tr>
<tr>
<td>NM_016772</td>
<td>Ech1</td>
<td>mnu00350</td>
<td>3</td>
<td>NM_021396</td>
<td>Pdcd1lg2</td>
<td>mnu04514</td>
<td>1</td>
</tr>
<tr>
<td>NM_010548</td>
<td>Il10</td>
<td>mnu04630</td>
<td>3</td>
<td>NM_013652</td>
<td>Ccl4</td>
<td>mnu04060</td>
<td>2</td>
</tr>
</tbody>
</table>
The fundamental strength of MicroPath stems from the implementation of a novel search strategy for the comparison of multiple gene expression profiles. Although there are a few software that cater for multiple gene expression comparison, there is currently no software that searches for common genes beyond simple number crunching methods of comparison (Table 5). Just because a direct comparison of a given number of datasets may not yield any common genes, it does not mean that the analysis should end here since there is a potential to identify common genes across n-1 profiles. MicroPath ensures that such genes are identified, which current software would overlook. When coupled with other important functions such as pattern extraction and pathway analysis, it becomes apparent that MicroPath would offer valuable assistance to biologists wanting to decipher their high throughput data.

Table 4. Results generated from pathway analysis showing the extent to which each pathway is affected by common genes from the intersection. The percentages reflect the proportion of common genes that contribute towards controlling the proportion of each pathway.

<table>
<thead>
<tr>
<th>Pathway ID</th>
<th>Pathway Name</th>
<th>GenBank Accession ID</th>
<th>Result from Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>mmu00071</td>
<td>Fatty Acid Metabolism</td>
<td>NM_007381</td>
<td>3.26% of genes contribute 8.45% role in pathway</td>
</tr>
<tr>
<td>mmu00280</td>
<td>Valine, leucine and isoleucine degradation</td>
<td>NM_007381</td>
<td>3.26% of genes contribute 2.73% role in pathway</td>
</tr>
<tr>
<td>mmu00410</td>
<td>Beta Alanine Metabolism</td>
<td>NM_007381</td>
<td>3.26% of genes contribute 7.14% role in pathway</td>
</tr>
<tr>
<td>mmu00640</td>
<td>Propanoate Metabolism</td>
<td>NM_007381</td>
<td>3.26% of genes contribute 5.88% role in pathway</td>
</tr>
<tr>
<td>mmu03320</td>
<td>PPAR Signalling Pathway</td>
<td>NM_007381</td>
<td>3.26% of genes contribute 1.92% role in pathway</td>
</tr>
<tr>
<td>mmu04514</td>
<td>Cell Adhesion Molecules</td>
<td>NM_007664 NM_008205 NM_013488 NM_021396</td>
<td>12.91% of genes contribute 6.84 % role in pathway</td>
</tr>
<tr>
<td>mmu04612</td>
<td>Antigen Processing &amp; Presentation</td>
<td>NM_013488</td>
<td>3.26% of genes contribute 2.44% role in pathway</td>
</tr>
<tr>
<td>mmu04640</td>
<td>Hematopoietic Cell Lineage</td>
<td>NM_013488</td>
<td>3.26% of genes contribute 0.76 % role in pathway</td>
</tr>
<tr>
<td>mmu04660</td>
<td>T Cell Receptor Signalling Pathway</td>
<td>NM_013488 NM_010548</td>
<td>6.45 % of genes contribute 3.33 % role in pathway</td>
</tr>
<tr>
<td>mmu04020</td>
<td>Calcium Signalling Pathway</td>
<td>NM_011696</td>
<td>3.26% of genes contribute 2.33 % role in pathway</td>
</tr>
<tr>
<td>mmu00350</td>
<td>Tyrosine Metabolism</td>
<td>NM_016772</td>
<td>3.26% of genes contribute 2.17 % role in pathway</td>
</tr>
<tr>
<td>mmu04060</td>
<td>Cytokine-cytokine receptor interaction</td>
<td>NM_010548 NM_013652</td>
<td>6.45 % of genes contribute 0.73 % role in pathway</td>
</tr>
<tr>
<td>mmu04630</td>
<td>JAK-STAT Signalling Pathway</td>
<td>NM_010548</td>
<td>3.26% of genes contribute 3.85 % role in pathway</td>
</tr>
<tr>
<td>mmu04670</td>
<td>Leukocyte Transendothelial Migration</td>
<td>NM_009510</td>
<td>3.26% of genes contribute 1.25 % role in pathway</td>
</tr>
<tr>
<td>mmu04810</td>
<td>Regulation of Actin Cytoskeleton</td>
<td>NM_009510</td>
<td>3.26% of genes contribute 1.47 % role in pathway</td>
</tr>
<tr>
<td>mmu04940</td>
<td>Type I Diabetes Mellitus</td>
<td>NM_008205</td>
<td>3.26% of genes contribute 4.35 % role in pathway</td>
</tr>
<tr>
<td>mmu00512</td>
<td>O-Glycan Biosynthesis</td>
<td>NM_013814</td>
<td>3.26% of genes contribute 10 % role in pathway</td>
</tr>
<tr>
<td>mmu04010</td>
<td>MAPK Signalling Pathway</td>
<td>NM_017777</td>
<td>3.26% of genes contribute 0.83 % role in pathway</td>
</tr>
<tr>
<td>mmu04620</td>
<td>Toll-Like Receptor Signalling Pathway</td>
<td>NM_019777 NM_013652</td>
<td>6.45% of genes contribute 1.32 % role in pathway</td>
</tr>
<tr>
<td>mmu04080</td>
<td>Neuroactive Ligand-Receptor Interaction</td>
<td>NM_010102</td>
<td>3.26% of genes contribute 1.15 % role in pathway</td>
</tr>
</tbody>
</table>

Table 5. Functional comparison of MicroPath to similar software packages and applications.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Suitable for high throughput data analysis</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Suitable for comparing multiple gene expression profiles</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Implementation of efficient algorithm to search for common genes from n-1 datasets</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Graphical representation of gene expression values from multiple datasets</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Pattern extraction from Graph data</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Construction of pathway maps</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Mapping gene expression data to pathway maps</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>User interactive software (S) or Database (D)</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>D</td>
<td>D</td>
<td>S</td>
</tr>
</tbody>
</table>
4. Conclusion

In this article, we have illustrated the potential benefits that can be derived from using MicroPath for the analysis of multiple gene expression profiles. Each function of the software has been developed to streamline the overall analysis pipeline, providing users with a walkthrough of how their data is biologically deciphered. Here, we have applied to our software, microarray datasets generated from different laboratories pertaining to the molecular mechanisms underlying immune tolerance. However, MicroPath is capable of analysing data for any given biological question, whether the datasets are taken from public repositories such as ArrayExpress or generated from in-house microarray experiments. We believe that its faculty to use both number crunching and permutations and combinations as the search strategy to identify the intersection of common genes, coupled with its function to extract gene expression patterns graphically and statistically makes this an attractive software for biologists to use. Finally, its ability to carry out live streaming of mapping genes to biological pathways makes it a useful tool for the automation of multiple gene expression analysis.

Availability and requirements

Project name: MicroPath
Project home page: www.1066technologies.co.uk/micropath
Operating system(s): MicroPath has been tested on Windows 2000, XP and Vista
Programming language: Visual Basic.Net, Perl
Other requirements: None
License: N/A
Any restrictions to use by non-academics: No

Acknowledgements

This study was supported by grants from the UK Medical Research Council (MRC) (Grant number: G0300520).

REFERENCES


Assessment of bone condition by acoustic emission technique: A review
Sharad Shrivastava, Ravi Prakash

ABSTRACT
The paper deals with the review of acoustic emission technique in biomedical field. The review is done with the aim to provide an overview of the use of AE technique in biomedical field, mainly concentrated on the AE behavior of bone under different loading conditions, its dependence on strain rate, in osteoporosis, monitoring the fracture healing process of bone. The overall conclusion from the review was that almost all the studies in bone indicated that the initial AE occurs only in the plastic region and just prior to yield. That means the use of AE technique for clinical application cannot be considered as a safe technique, but the early occurrence of AE events from callus promises the application of AE technique for monitoring the fracture healing process. The negligible effect of soft tissues on AE response of bone promises AE to become a non-invasive method for assessment of bone condition.

Keywords: Acoustic Emission; Assessment; Strain Rate; Callus; Fracture Healing; Osteoporosis

1. INTRODUCTION
Bone is primary structural element of human body. The anatomy of human beings is quite well known but the strength and mechanical properties of bones have not been investigated thoroughly. The 206 named bones of skeleton constitutes 18% of the adult human body weight, only skin and fat (25%) and muscles (43%) being greater [1]. In biological terms bone is described as a connective tissue and in mechanical terms bone is a composite material with several distinct solid and fluid phases. The mechanical properties of bone have been more extensively investigated than those of any other biological tissue materials. Although our understanding of the mechanical properties and fracture behavior of bone is continuously improving, as yet it is far from complete. As pointed by Hayes, W. C. [2] while fundamental research is needed on many aspects of the mechanical response of the bone, applications of the techniques of analytical and experimental mechanics in this area are made complicated by the fact that bone is highly complex living material.

The initial work in the field of bone biomechanics can be traced back to the 17th century when “attempts to express biological findings in physical terms” [3] were made by the scholars at that time. This philosophical background of the aspect was intensified in the age of determinism which lasted until the middle of the 19th century. One of the main aspects of the research work in that time was to relate the architecture of bone and its mechanical functions. In the year 1832, Bourgery, J. M. [4] in his work on anatomy raised the question of relation between architecture and mechanical functions of bone. In his book on osteology, Ward, F. O. [5] compared the proximal end of the human femur with a crane and he mentioned the compressive and tensile stresses evoked in the bone by loading. In the year 1867, a more detailed analysis of the structure of cancellous bone and its mathematical significance was given by Meyer, G. H. [6] in association with the famous mathematician Cullmann.

The industrial revolution took place in the second half of the 19th century. It had an impact on the research works in the bone also. New developments were made in the field of material testing and the new methods were developed for mechanical measurements. For a while, these methods were used to determine the in vitro mechanical properties of bone. The bones were tested under various loading conditions and the ultimate strength of bone was determined by many investigators [7,8,9,10,11,12,13,14,15,16,17,18]. Mc Elhaney, J. H. [19] from his study on the strain rate dependence of the mechanical properties of bone showed that both the compressive strength and modulus of longitudinally oriented compact bone specimens were significantly increased by increasing the strain rate. A critical strain rate for bone has been
claimed in compression [19], torsion [20] and tension [21]. However Wright, T. M. and Hayes, W. C. [22] found no critical strain rate in tensile tests of bovine bone over a wide strain-rate range.

In the last few decades attempts were made to use the newly developed/improved non-destructive testing techniques to find the mechanical properties of in vivo and in vitro bones. Those include finding the elastic constants using ultrasonic techniques [23,24,25,26], finding the mechanical strength of bone specimens by X-ray computed tomography, etc.

Assessment of in vivo bone condition is one of the research areas, which have attracted many biomedical engineers and clinical orthopaedicians in recent times. Presently the radiological examination is widely used for the assessment of in vivo bone condition [27,28]. In some clinical problems such as diagnosis of the point of clinical union of fracture, the manual assessment of stability is also used along with the radiological examination. However for many applications the radiographic technique was found to be suffered from low sensitivity. For instance, for the evaluation of osteoporosis it requires a minimum loss of 30% or more of bone mineral content before an unequivocal roentgen logical diagnosis can be made [29].

Monitoring the fracture healing process is another area where the currently used techniques failed to give satisfactory results. Uncertainty regarding the significance of the radiographic and clinical findings may result in unnecessarily long immobilization periods which can produce discomfort and inconvenience for the patients, as well as possible joint stiffness and even permanent loss of motion especially in the elderly.

In certain long bone shaft fractures the healing process is modified by the method of treatment so that the clinical assessment of mechanical integrity is impossible and the interpretation of radiographs may be difficult.

Diaphyseal fractures treated by “rigid” internal fixation always demonstrated this problem, since the fracture cannot be tested mechanically and external callus formation is not seen on radiographs, methods are needed to assess the mechanical integrity of fracture healing in such circumstances, or an unreliable and unsafe rehabilitation programme may be prescribed.

Mechanical impedance, natural frequency, vibration analysis, stress wave propagation, ultrasonic measurements, impact response technique, electrical potential measurements and mechanical tissue response analysis are some of the techniques, which have been attempted by different investigators for the assessment of in vivo bone condition in the past [30,31,32,33,34,35,36,37,38,39,40]. However, in all these studies, the intervening soft tissues, whose quantity and quality changes with individual to individual, affected the results. Furthermore some methods were not really non-invasive in nature and some were not practicable for widespread clinical use because of low reliability and complicated instrumentation.

The structure of bone is very much similar to engineering composite materials and is therefore advantageous to use a non-destructive testing technique, which has already proved it usefulness in the field of composite materials testing. Acoustic emission (AE) technique has been used very successfully for the non-destructive evaluation of composites. The relationship between AE response and mechanical behavior in composite materials has been extensively studied in the past [41]. This paper deals with the review of AE technique which is to be used for bone assessment. The review has been broadly classified as follows (Figure 1).

2. ACOUSTIC EMISSION TECHNIQUE

The AE technique is the sound produced by materials as they fail. A familiar example is the audible cracking

![Acoustic Emission Technique](image)

**Figure 1.** Broad classification of the review study.
noise from wood. Almost all engineering materials generate acoustic emissions but unlike wood, the sound is too faint to be heard without sensitive electronic monitors. Acoustic emission waves can be detected by means of remote piezoelectric sensors and their source can be located by timing the wave arrival at several sensors. Thus AE provides a unique method of recognizing when and where deformation is taking place as a structure is stressed.

The first systematic investigations of AE phenomenon was made in 1950 by Kaiser, J. [42] at the technical university of Munich. In his investigations, the noise emitted by the deformation of materials was examined by means of electronic equipment capable of detecting inaudible ultrasonic signals. Kaiser, while working with polycrystalline specimens concluded that acoustic vibrations originate in grain boundary interfaces and was believed to be associated with the interaction induced between interfaces by applied stresses. He noted that, for a given material, characteristics spectra of frequency and amplitude existed. One of the important observations made in his study was that irreversible processes were involved with AE phenomenon; an effect later came to be known as Kaiser Effect. The universality of the AE phenomenon, as recognized by Kaiser, leads to a very wide range of applicability. AE has been recorded from hundreds of materials—metals, composites, ceramics, plastics, glasses, building materials, biological materials in vitro and in vivo as well as from multi material structures and joints between different materials [43]. Compared to other NDT techniques which rely on extraneous energy for the illumination of defect; AE enjoys the unique feature that the defect makes its own signal. This leads to a natural complementarity between AE and other methods.

3. ACOUSTIC EMISSION DETECTION AND SIGNAL PROCESSING

Figure 2 [44] shows the method of detection of acoustic emission events by remote piezoelectric transducers. Here, an AE source generates an expanding spherical wave packet losing intensity at a rate of $r^{-2}$. When this wave reaches the body boundary, a surface wave packet is created, either Rayleigh or Lamb wave type depending on the thickness. This method is mainly used for flaw monitoring in inaccessible areas. The individual signal has a short duration at the source and a corresponding broad spectrum which typically extends from zero frequency to many megahertz. The form of the signal at the point of detection is a damped oscillation, developed in the structure according to known principles of acoustic wave propagation. Figure 3 [44] shows the signal waveform of one acoustic emission event and different parameters normally measured to characterize the acoustic emission source.

4. ACOUSTIC EMISSION IN BONE

Hanagud, S., et al. [45] using bovine femora first demonstrated detectable acoustic emissions from bone. His work made the way for other investigators to use the AE technique for characterization of bone and also to explore the possibilities of using it as a tool for clinical orthopaedicians to detect bone abnormalities [46, 47, 48, 49, 50]. Knet-s, I. V., et al. [46] has shown that the character of the fracture surface depends on the orientation of the load relative to the direction of the osteons, the rate of loading, and the geometrical shape of the actual sample. They concluded that the most promising approach in testing the internal state of a bone is acoustic emission, sometimes also known as the method of stress-wave emission. This approach involves recording of deformation noise in the material due to the development and further propagation of structural defects. These defects may include dislocations or cracks appearing in the course of loading. This study was the first work to visualize the degree of micro cracks development in bone tissue subjected to longitudinal extension. This experimental work was only visualized for longitudinal loading and the deformation rate considered was also very low (1 mm/min). The study was not conducted for higher strain rates. In another investigation Hanagud, S., et al. [50] conducted AE tests on carefully prepared bone specimens subjected to bending loads. Their specimen included femur from cattle and cadavers. They compared the AE patterns from 60 perfect and defective specimens. The result clearly indicated that the development of an effective early diagnostic tool for osteoporosis was possible by using AE technique.

Thomas, R. A., et al. [51] studied the acoustic emissions from fresh bovine femora and its clinical applications. They employed a more sophisticated set up of AE technique by including both amplitude and pulse width distribution to investigate whole fresh bovine femora which were loaded by compression and bending. They found that both the amplitude distribution and pulse width distribution results of fresh bone had clearly shown characteristic spectra which could be used for the early detection of bone abnormalities such as fracture and osteoporosis.
Yoon, H. S., et al. [52] developed a new AE technique for applications to human and animals both non-invasively and non-traumatically. Bones from several different species of animals and different kinds in the same species were tested to obtain AE parameters. Their results indicated that the AE amplitude distributions of all the bones are similar, somewhat independent of the species of animals and kind in the same species and however different from those of those materials such as metals, ceramics and plastics. The technique was found useful for the diagnosis of micro fractures, such as stress fractures in the tibia of runners, which were not detectable by conventional X-ray technique until they began to heal. Moreover conventional techniques would require introducing some additional stresses in to the part of the body under examination, which also introduces additional trauma to the patient. In their technique the low intensity ultrasonic pulses were injected through an AE transducer, instead of applying loads to the bone under tests. The loading as pulses reduced the introduction of trauma to the live subject. Another receiving AE transducer was used to collect five types of useful AE data: per-event distribution of counts, peak amplitude, energy and pulse duration, and cumulative counts vs time.

Netz, P. [53] monitored the AE response of canine femora in torsion at 6 degrees per second. His work demonstrated that the AE events occur in the non-linear plastic portion of the load deflection curve. Wright, T. M., et al. [54] monitored the permanent deformation of compact bone using AE technique. Uniaxial tension tests were performed on standardized specimens of bovine harvesian bone to examine the contributions of mineral and collagen to permanent deformation in bone and to monitor the damage mechanisms occurring in permanent deformation using AE technique. Their results were consistent with a two-phase model for bone in which the mineral behaves as an elastic-perfectly plastic material when bound to the collagen fiber matrix. The AE events occurred just prior to the yield point and continued during yielding. Significant AE counts occurred again just prior to fracture. No emissions occurred in the elastic region and few occurred in the major portion of plastic region between yield and fracture. To monitor micro cracks in the specimen they used AE and plotted graphs. Figures 4, 5 [54] show stress vs. strain and cumulative acoustic emission counts vs. strain curves for one of the control specimens and decalcified specimens. These graphs indicate the similarity between the acoustic emission data of the bones prior to fracture. Figure 6 [54] shows stress strain plots based on the mean values from Table 1 [54]. The limitation of their work lies in the hypothesis that the mineral only exhibit elastic-perfectly plastic behavior in conjunction with collagen. They conducted experiments on control, decalcified and deprotenised groups of specimens with the same hypothesis. The fact is that the deprotenised groups of specimens behave in brittle manner. Hence they suggested further studies to be undertaken to examine the contributions of mineral and collagen for permanent deformation in the bone. The two phase model used could be used to study the quasi-static tensile behavior of compact bone but more work could have been carried out for higher strain rates responses before coming to any conclusion.
Figure 6. Stress-strain curves for the three test groups constructed from the mean values in Table 1. Error bars are shown for ultimate stress values (source: 54).

Table 1. Mechanical properties of decalcified and partially deproteinized bovine bones (source: 54).

<table>
<thead>
<tr>
<th></th>
<th>control</th>
<th>decalcified</th>
<th>Deproteinized</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of specimens</td>
<td>7</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Yield stress</td>
<td>118(9.8)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yield strain</td>
<td>0.544(0.150)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ultimate stress</td>
<td>128(15.6)</td>
<td>34(7.5)</td>
<td>71(11.5)</td>
</tr>
<tr>
<td>Ultimate strain</td>
<td>2.02(0.924)</td>
<td>9.247(1.524)</td>
<td>0.956(0.342)</td>
</tr>
<tr>
<td>Elastic modulus</td>
<td>20.6(2.76)</td>
<td>0.37(0.05)</td>
<td>11.3(3.15)</td>
</tr>
<tr>
<td>Plastic modulus</td>
<td>0.66(0.357)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

It is an established fact that the ultimate tensile strength of bone is dependent on the applied strain rate. Based on these Fisher, R. A., et al. [55] studied the effect of using two different strain rates on the AE in bones. In their work, bovine cortical bone was milled into standard tensile specimens which were tested at two different strain rates while being monitored with AE equipment. They found that the amplitude distribution of the AE events in bone is dependent on strain rate. Greater number of events occurred with the slower strain rate but the events were of lower amplitude than those emitted during the more rapid strain rate. Here also the initial AE occurred well in to the plastic region of the stress-strain curve near the point of fracture of the tensile specimens. It was evident from the study that if acoustic emission technology is to be utilized clinically for the assessment of fracture healing; careful selection of rate of loading would be necessary. Furthermore the study indicated that acoustic emission response was different at different strain rates. As the emissions did not occur until failure was imminent, it indicates that Acoustic emission technology is not suitable for evaluating the integrity of bone. The limitation included that the specimens taken were of very large sizes to minimize the stress concentrations effects of normal bone architecture. They did their work at only two different strain rates (0.0001/s and 0.01/s); therefore no conclusion could be made for broad range of strain rates.

Later on Nicholls, P. J., et al. [56] studied the AE properties of callus. In their work, rabbits with 45 degree midshaft oblique osteotomies were strained in shear while monitoring for AE events. Each fracture remained essentially quiet until over 50% of load to failure had been applied. They suggested that since callus formation during fracture healing takes important role in the healing process, the AE from callus may have clinical applications. The limitation of the study lies in the test method used for evaluating bone. As bone is a non-homogeneous substance, it is very difficult to evaluate with instruments which have been designed for homogeneous substances. A test method should be so designed that eliminates background noise, such as slippage of specimen in the grips and motion of the transducers on the bone surface. This hampers reproducibility of acoustic emission patterns. In most of the AE studies of bone the bone has been tested without the surrounding soft tissues. But in the case of clinical applications of AE, one cannot separate the bone from soft tissues and hence the tests should be performed with soft tissues. Hanagud, S., et al. [57] studied these phenomena. They used freshly dissected rabbit tibia and femur with soft tissues. Tests were conducted through bending load. They found that the soft tissues of 2 to 9mm thickness did not affect the bone’s AE response.

A study of bone-tissue samples by Martens, M. [58] used acoustic emission for to study the mechanical behavior of femoral bones in bending loading. Ono, K. [59], provided an insight about the fundamental theories and equations related to acoustic emission. Stromsoe, K., et al. [60], worked on bending strength of femur using non-invasive bone mineral assessment.

The work done till this time demonstrated that the safe use of AE technique for the non-destructive testing of bone is impossible because the AE events occurred only after plastic deformation occurred.

Lentle, B. C. [61], University of British Columbia used acoustic emission to monitor osteoporosis. He devised a method for in vivo diagnosis of patients using AE technique, which could also predict the severity of osteoporosis.

Table 2. The predictive capability of acoustic emissions expressed in terms of the specimen’s fatigue life (source: 71).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specimen 1</td>
<td>55</td>
<td>26363</td>
<td>22580</td>
<td>85 %</td>
</tr>
<tr>
<td>Specimen 2</td>
<td>61</td>
<td>35020</td>
<td>33382</td>
<td>95 %</td>
</tr>
<tr>
<td>Specimen 3</td>
<td>66</td>
<td>4737</td>
<td>2989</td>
<td>63 %</td>
</tr>
<tr>
<td>Specimen 4</td>
<td>71</td>
<td>600</td>
<td>402</td>
<td>67 %</td>
</tr>
</tbody>
</table>
A statistically significant effect of time on these mechanical properties was detected. Within a row, values with differing letters are significantly different from each other (P<0.05) (Source: 65).

Table 3. Ash content was calculated by (ash density/apparent density Â· 100). A statistically significant effect of time on these mechanical properties was detected. Within a row, values with differing letters are significantly different from each other letters (P<0.05) (Source: 65).

<table>
<thead>
<tr>
<th>Mechanical properties</th>
<th>4(n=8)</th>
<th>6(n=9)</th>
<th>8(n=10)</th>
<th>12(n=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apparent density (g/cm³)</td>
<td>0.42±0.07a</td>
<td>0.67±0.07b</td>
<td>1.2±0.1c</td>
<td>1.2±0.08c</td>
</tr>
<tr>
<td>Ash density (g/cm³)</td>
<td>0.14±0.04a</td>
<td>0.40±0.06b</td>
<td>0.82±0.08c</td>
<td>0.88±0.06c</td>
</tr>
<tr>
<td>Ash content (%)</td>
<td>33±4a</td>
<td>59±5b</td>
<td>71±1b,c</td>
<td>73±1c</td>
</tr>
</tbody>
</table>

Acoustic emission was being used to predict changes in mechanical properties due to fatigue [62,63,64]. Watanabe, Y., et al. [65] used AE technique to predict mechanical properties of fractures. Experimentally produced fractures of femur in rats were tested in tension and in torsion at 4, 6, 8 and 12 weeks after fracture. AE signals were monitored during these mechanical tests. The values for load and torque at the initiation of the AE signal were defined as new mechanical parameters. Tensile strength, tensile stiffness, and torsional stiffness were found to increase with time. They focused on how AE signals can help a surgeon to remove the external fixators in the sense that AE signals can be used to monitor healing of bones. Table 3 [65] indicates a statistically significant effect of time on these mechanical properties. Table 4 [65] indicates the calculated ash content and the statistically significant effect of time on the mechanical properties of bone. The data obtained by them were compared to the original values and were found out to be almost the same. The study was a first step towards the establishment of AE testing as a means of predicting the callus strength. The two parameters exhibited strong and positive linear correlation with tensile strength and torque. The linear correlations suggested that it may be possible to use AE technology to evaluate fracture healing process, following osteotomy surgery. There were still many issues which needed to be resolved to make it clinically viable.

Kevin S. C. K., et al. [66] developed an acoustical technique for the measurement of structural symmetry of hip joints. Since, these techniques depend very much on the intensity and quality of sounds emitted from the joints under investigation. They developed an acoustical technique for the measurement of relative acoustic transmission across both hips of the test subjects while they were subjected to an external vibratory force applied at the sacrum. The merit of this approach was that it allows direct comparison of the sound signals transmitted across both hips regardless of the measure of the input vibratory force. Simultaneously, other acoustic techniques like scanning acoustic microscopy [67,68] acoustic mapping [69,70] was being used to predict and study mechanical properties of tissues and bone.

Ozan A. [71] worked on a hypothesis that an increase in micro damage activity during repeated loading of bone will signal the approaching stress fracture. Interception with the training regime prior to the incidence of the fracture as signaled by acoustic emissions would reduce the time necessary for recuperation. Acoustic emission was used for real time monitoring of micro cracks. They used acoustic emission technique to predict the failure of cortical bone. Table 2 [71] indicates the predictive capability of acoustic emissions expressed in terms of the specimen’s fatigue life.

Information was collected on all acoustic events, regardless of whether they originated from micro damage or somewhere else and then signals originating from the micro damage were isolated. The rest of the irrelevant signals were filtered out based on their average frequency, duration, amplitude, and intensity. With the non-micro damage signals removed from the data, we were able to determine the number acoustic events related to bone damage as well as the time at which they occurred. Specially designed software is yet to be developed which will segregate the zones of micro damage. Fracture healing and prediction of healing time of fractures were increasingly being studied. A review by Browne, M., et al. [72] on acoustic emission’s capability to monitor bone degradation and bone fatigue provided us information with latest developments in this field.

In 2004, Franke, R. P., et al. [73] used acoustic emission for in vivo diagnosis of the knee joint. For the as-
assessment of the tribological knee function and by the probability of fracture of the femur an adapted Acoustic Emission Measurement System named Bone Diagnostic System (BONDIAS) was developed. This system makes the in vivo analysis of the medical status possible. Different mechanisms of cracking were accompanied by different acoustic emission from human femora as shown in literature. An acoustic emission signal typical of crack initiation is shown in Figure 7 [73]. This Figure is indicative of the acoustic emission from healthy knee joint cartilage after a sudden change from a two leg stand to a one leg stand. It is characterized by a very short rise time and an exponential decrease of the amplitudes. From the medical point of view such mechanical loads are regarded as non-destructive although there is already crack initiation in the interface of the compact and the trabecular system of the bone. These micro cracks seem to be essential for the physiological bone remodeling. For the description of the development of bone strength over time it is necessary to assess both the threshold of crack initiation and the conditions for crack propagation. The sudden change in amplitude indicates high thickness of the cartilage layer. There are several advantages of the diagnostic procedure by AE when compared with established conventional methods:

1) No pain is caused by this procedure.
2) This procedure is non-destructive. Mechanical load even beyond the crack initiation threshold are typical of day to day life and necessary for the physiological bone remodeling to avoid the degeneration of the bone and joint system.
3) There is no health burden through ionizing radiation as is unavoidable with X-ray examination and CT.
4) There is no danger of infection since this is a non-invasive examination.
5) The time required for the assessment of the acoustic emission behavior and analyses of data are of the order of seconds to minutes.

6) The expenses for the AE measurement system are small compared to X-ray systems.
7) The costs per examination including a detailed diagnosis are well below costs of other diagnostic procedures and there is no danger of infection leading to further costs, as happens with invasive methods, e.g. endoscopic examinations.
8) Diagnostic (Real time) monitoring of bone and joint training of sports professionals becomes possible.

The disadvantage of the measurement system suggested was that the physician will be left with the bundle of data and the task to evaluate the AE.

Tatarinov, A., et al. [74] proposed multiple acoustic wave method for assessing long bones. The method was based on measurement of ultrasound velocity at different ratio of wavelength to the bone thickness and taking into account both bulk and guided waves. They assessed the changes in both the material properties related to porosity and mineralization as well as the cortical thickness influenced by resorption from inner layers, which are equally important in diagnosis of osteoporosis and osteopenia. More in vivo studies on animals and human volunteers has to be carried out before the proposed method could be made clinically usable. The advantage of the method proposed was that it allowed assessment of changes in both the material properties related to porosity and mineralization as well as cortical thickness influenced by resorption from inner layers, which are equally important in diagnosis of osteoporosis and other bone osteopenia. The method could also be used for diagnosis of bone condition if the contribution of soft tissues and topographical heterogeneity in real bones are considered. The method had a potential for better detection of early stage of osteoporosis in long bones. Singh, V. R. [75] reviewed an acoustic imaging technique known as acoustic stress wave propagation technique which was used for bone examination. The technique was developed with the view to solve the problems encountered with the conventional technique like X-rays. As bone is a heterogeneous, complex and fibrous tissue, determination of very small abnormalities viz. shape and size of bone defects, is not usually possible by means of conventional X-ray technique.

In 2006, Azra Alizad, et al. [76] studied the change in resonant frequencies of a bone due to change in its physical properties caused due to a fracture. Experiments were conducted on excised rat femurs and resonance frequencies of intact, fractured, and bonded (simulating healed) bones were measured. These experiments demonstrated that changes in the resonance frequency indicated bone fracture and healing. The fractured bone exhibits a lower resonance frequency than the intact bone, and the resonance frequency of the bonded bone approaches that of the intact bone. The graphs are indicative of the result (Figures 8,9) [76], that the frequency re-
sponse of a cut femur is less than the intact femur. The proposed method may be used as a remote and non-invasive tool for monitoring bone fracture and healing process, and the use of focused ultrasound enables one to selectively evaluate individual bones. The proposed method offers several advantages over vibrational methods using external mechanical excitation. The ultrasound can be applied remotely and directly to the bone under test, thus avoiding interference of overlaying muscle or other tissues on force distribution. Furthermore, in contrast to traditional methods in which it is difficult to target small bones and to access them, the proposed method allows application of excitation force directly and selectively to the intended bone. The acoustic method for measuring

![Figure 8. Frequency response of the intact femur A. The plots show the motion of the intact femur vs frequency. These Fig plots indicate peaks at 925 Hz, 4.2 kHz, and 8.1 kHz. Peaks of motion below 700 Hz were explored and found not to be related to the femur. Top left: Driven and measured at the end of the femur at frequency range of 100 Hz to 1000 Hz. Top right: Driven and measured at the end of the bone at 1 kHz–10 kHz. Bottom left: Driven and measured at bone midpoint at 100 Hz–1000 Hz. Bottom right: Driven and measured at bone midpoint at 1 kHz–10 kHz (source: 76).](image)

![Figure 9. Frequency response of the cut femur (source: 76).](image)
bone response is suitable for in vivo applications as long as one take the frequency response of the surrounding structures into account. An advantage of acoustic motion detection method is that it does not require a direct path to the bone because the acoustic emission produced by the bone travels easily in every direction, thus the location of the hydrophone is not critical. Further investigations are needed to demonstrate the applicability of the proposed method for evaluation of bone quality in human body. The boundary conditions of bone to body must also be considered before applying the proposed method.

In 2008, Dipan Bose, et al. [77] studied the effect of valgus bending and shear loading on knee joint. They used acoustic sensors to determine the failure timing of soft tissues attached to femur and tibia. The failure timing was determined based on the knee injury mechanism due to valgus loading. At the estimated time injury, the corresponding values of \( \alpha_{\text{valgus fail}}, \alpha_{\text{shear fail}}, M_{\text{valgus fail}}, \) and \( V_{\text{shear fail}} \) were designated as the failure parameters of the knee. Numerical methods with accurate geometric and material properties could be implemented to simulate and further extend the injury threshold to alternate loading methods. Elmar K. Tschegg, et al. [78] did stiffness analysis of tibia implant system under cyclic loading. He used a bio-mechanical system integrated with acoustic emission sensors at the screw head. 3 sequences of loading were used to determine when the locking screws break. Data was obtained from the acoustic sensors onto a data acquisition board and were processed using a acoustic emission software. This acquired data was used to determine as to which screw is bearing the load and as to when does the screw break.

5. CONCLUSIONS

- Many investigations carried out in the field “Assessment of bone condition” are mainly in vitro studies. For any method which is going to be used in clinical practice, a thorough experimental study with animals and/or a clinical study with human volunteers are very much essential.
- The AE is very much dependent on the strain rate.
- AE technique is highly sensitive to specimen damage and cracks and detects them even before visual detection.
- Diagnostic (Real time) monitoring of Bone is possible.
- It is non-destructive and helps us to predict the time length of the healing process.
- It has no harmful effects unlike X-rays, which have radiation effects on patients.
- The emissions from the callus during fracture healing, gives its possibility to be clinically used.

6. FUTURE RESEARCH DIRECTIONS

- Many researchers have used this technique for in vitro as well as in vivo characterization of bone. However, the clinical application of the technique was not fully investigated. The time of occurrence of initial AE and the AE response of bone under different loading conditions at different strain rates are not well established. So focus can be on finding the exact time of occurrence of initial AE with respect to the stress/strain curve and the AE behavior under a suitable loading condition at different strain rates.
- In the case of in vivo studies, the AE response of callus is not thoroughly investigated. Also it is necessary to conduct an experimental study with laboratory animals/volunteers or patients to prove the clinical usage of AE.

7. ACKNOWLEDGEMENT

The authors wish to acknowledge Mr. Sudeep Mohapatra & Mr. Kaushik V. for their help. Financial assistance provided by Department of Science and Technology, Government of India is also gratefully acknowledged.

REFERENCES


Descriptively probabilistic relationship between mutated primary structure of von Hippel-Lindau protein and its clinical outcome

Shao-Min Yan, Guang Wu

ABSTRACT
In this study, we use the cross-impact analysis to build a descriptively probabilistic relationship between mutant von Hippel-Lindau protein and its clinical outcome after quantifying mutant von Hippel-Lindau proteins with the amino-acid distribution probability, then we use the Bayesian equation to determine the probability that the von Hippel-Lindau disease occurs under a mutation, and finally we attempt to distinguish the classifications of clinical outcomes as well as the endocrine and nonendocrine neoplasia induced by mutations of von Hippel-Lindau protein. The results show that a patient has 9/10 chance of being von Hippel-Lindau disease when a new mutation occurs in von Hippel-Lindau protein, the possible distinguishing of classifications of clinical outcomes using modeling, and the explanation of the endocrine and nonendocrine neoplasia in modeling view.

Keywords: Amino Acid; Bayes’ Law; Cross-Impact Analysis; Distribution Probability; Mutation; Von Hippel-Lindau Disease

1. INTRODUCTION
Perhaps, the first step to study the genotype-phenotype relationship is to determine a protein in relation to a disease, and the second step would be to build a quantitative relationship between mutant protein and its clinical outcome. Then we may be in the position to predict the clinical outcome based on such a quantitative relationship, even to predict new functions led by new mutations.

Thus, we need the methods, which can quantify a protein sequence as a numeric sequence in order to build a quantitative relationship. In fact, we have various ways to quantify a protein sequence, for example, to use the physicochemical property of amino acid to quantify a protein sequence [1].

Since 1999, we have developed three approaches to quantify each amino acid in a protein as well as a whole protein (for reviews, see [2,3,4]), and our quantifications indeed differ before and after mutation, thus it is possible to use our approaches to build a quantitative relationship between changed primary structure and changed function of protein.

In 1911 and 1926, von Hippel and Lindau described the von Hippel-Lindau disease [5,6], later on Melmon and Rosen established the notion of the von Hippel-Lindau disease [7], which is an autosomal dominant disorder characterized by cerebellar, spinal cord, and retinal hemangioblastomas; cysts of the kidney, pancreas, liver, and epididymis; and has an increased frequency of renal cancer (renal cell carcinoma or hypernephroma), pancreatic cancer, and pheochromocytoma [8,9,10]. The von Hippel-Lindau disease has a birth incidence of about 1 in 36000 and about 20% of cases arise as de novo mutations without a family history [11,12].

The von Hippel-Lindau disease tumor suppressor gene was identified in 1993 [13], of which mutations are the major cause for developing the von Hippel-Lindau disease. Pathologically relevant is inactivation of the von Hippel-Lindau gene and subsequent loss of the function of the von Hippel-Lindau protein, and Elongin B, C complex [14,15]. The dysfunction of the ubiquitination of hypoxia-inducible factors is an important step in the development of various tumors [15,16,17,18,19]. Also, a recent study elucidated the role of NGF/JunB/ EglN3-related pathways in developmental apoptosis linking to tumourigenesis [20].

Clinically the von Hippel-Lindau disease is classified into two types: type I without pheochromocytoma and type II with pheochromocytoma [10,17]. On the other hand, more than 300 different von Hippel-Lindau muta-
tions have been described at DNA level [21, 22, 23, 24], and more than 100 at protein level. It would be great helpful if we can build a quantitative relationship between von Hippel-Lindau protein mutation and von Hippel-Lindau disease status, that is, the relationship between mutant protein and its clinical outcome.

In this study, we build a descriptively quantitative relationship between changed primary structure of mutated von Hippel-Lindau protein and the classification of its clinical outcome, distinguish the classifications of clinical outcomes as well as the endocrine and nonendocrine neoplasia induced by mutations of von Hippel-Lindau protein.

2. MATERIALS AND METHODS

2.1. Data

The human von Hippel-Lindau disease tumor suppressor with total 132 mutations (accession number P40337; December 4, 2007; Entry version 91) is obtained from UniProtKB/Swiss-Prot entry [25]. Among them, 123 are missense point mutations, 7 deletions and 3 insertions.

2.2. Amino-Acid Distribution Probability

Among three approaches developed by us, the amino-acid distribution probability is mainly related to the positions of amino acids along the protein, which is suitable for mutation analysis, and we have used this approach in a number of our previous studies {2, 3, 4, 6, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44}. The quantification is developed along such a thought, for example, how do two amino acids distribute along a protein sequence? Our intuition may suggest that there would be one amino acid in the first half of the sequence and another one in the second half. In fact, there are only three possible distributions, 1) both amino acids are in the first half, 2) one amino acid is in each half and 3) both amino acids are in the second half. Thus, each distribution has the probability of 1/3. If we do not distinguish either the first half or second half but are simply interested in whether both amino acids are in both halves or in any half, there will be the probability of 1/2 for each distribution.

If we are interested in the distribution probability of three amino acids in a protein, we naturally imagine to grouping the protein into three partitions, and our intuition may suggest that each partition contains an amino acid. If we do not distinguish the first, second and third partition, actually there are totally three types of distributions, i.e. 1) each amino acid is in each partition, 2) two amino acids are in a partition and an amino acid is in another partition, and 3) three amino acids are in a partition.

In this situation, the distribution probability can be calculated according to the statistical mechanics, which classifies the distribution of elementary particles in energy states according to three assumptions of whether distinguishing each particle and energy state, i.e. Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein assumptions [45]. We usually use the Maxwell-Boltzmann assumption for computing amino-acid distribution probability, which is equal to \( \frac{r!}{q_1!q_2! \cdots q_n!} \times \prod_{k=1}^{n} q_k^{r_k} \), \( r \) is the number of amino acids, \( n \) is the number of partitions, \( r_k \) is the number of amino acids in the \( n \)-th partition, \( q_k \) is the number of partitions with the same number of amino acids, and \( ! \) is the factorial function.

Thus, the distribution probabilities are different for these three types of distributions of three amino acids, say, 0.2222 for 1), 0.6667 for 2) and 0.1111 for 3).

Clearly the protein can only adopt one type of distribution for these three amino acids, which is the actual distribution probability.

For four amino acids, there are five distributions, 1) each partition contains an amino acid, 2) a partition contains two amino acids and two partitions contain an amino acid each, 3) two partitions contain two amino acids each, 4) a partition contains an amino acid and a partition contains three amino acids, and 5) a partition contains four amino acids. Their distribution probabilities are 0.0938 for 1), 0.5625 for 2), 0.1406 for 3), 0.1875 for 4), and 0.0156 for 5). Furthermore, there are seven distributions for five amino acids, 11 distributions for six amino acids, 15 distributions for seven amino acids, and so on.

2.3. Quantification of Wild-Type von Hippel-Lindau Protein

Table 1. Amino acids, their composition and distribution probability in wild-type human von Hippel-Lindau protein. (A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; E, glutamic acid; Q, glutamine; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine.)

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Number</th>
<th>Distribution probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>0.0476</td>
</tr>
<tr>
<td>R</td>
<td>20</td>
<td>0.0067</td>
</tr>
<tr>
<td>N</td>
<td>9</td>
<td>0.1770</td>
</tr>
<tr>
<td>D</td>
<td>11</td>
<td>0.1077</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.5000</td>
</tr>
<tr>
<td>E</td>
<td>30</td>
<td>0.0001</td>
</tr>
<tr>
<td>Q</td>
<td>8</td>
<td>0.0673</td>
</tr>
<tr>
<td>G</td>
<td>18</td>
<td>0.0389</td>
</tr>
<tr>
<td>H</td>
<td>5</td>
<td>0.0640</td>
</tr>
<tr>
<td>I</td>
<td>6</td>
<td>0.1543</td>
</tr>
<tr>
<td>L</td>
<td>20</td>
<td>0.0422</td>
</tr>
<tr>
<td>K</td>
<td>3</td>
<td>0.1111</td>
</tr>
<tr>
<td>M</td>
<td>3</td>
<td>0.6667</td>
</tr>
<tr>
<td>F</td>
<td>5</td>
<td>0.2880</td>
</tr>
<tr>
<td>P</td>
<td>19</td>
<td>0.0319</td>
</tr>
<tr>
<td>S</td>
<td>11</td>
<td>0.0404</td>
</tr>
<tr>
<td>T</td>
<td>7</td>
<td>0.2142</td>
</tr>
<tr>
<td>W</td>
<td>3</td>
<td>0.6667</td>
</tr>
<tr>
<td>Y</td>
<td>6</td>
<td>0.2315</td>
</tr>
<tr>
<td>V</td>
<td>17</td>
<td>0.1280</td>
</tr>
</tbody>
</table>
With respect to the wild-type von Hippel-Lindau protein, for example, there are eight glutamines “Q” in von Hippel-Lindau protein (Table 1). We may ask how these eight Qs distribute along the von Hippel-Lindau protein? According to the problem of the occupancy of subpopulations and partitions [45], the simple way to answer this question is to imagine that we would divide the von Hippel-Lindau protein into eight equal partitions, and each partition has about 27 amino acids (213/8 = 26.625) because the von Hippel-Lindau protein is composed of 213 amino acids, then there would be 22 configurations for all the possible distributions of eight Qs (Table 2).

Here, we calculate two distribution probabilities in Table 2 as example according to the above equation. For eight Qs equally distribute in each partition (the second row in Table 2), we have \( q_0 = 2, q_1 = 1, q_2 = 0, q_3 = 1, q_4 = 0, q_5 = 0, q_6 = 0, q_7 = 0, q_8 = 0; \) and \( r_1 = 0, r_2 = 0, r_3 = 1, r_4 = 1, r_5 = 1, r_6 = 1, r_7 = 1, r_8 = 3, \) that is,

\[
\frac{8!}{40320 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1} \times \frac{8!}{40320 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1} \times 8^{-8} = \frac{1}{16777216} = 0.002403
\]

Clearly, the von Hippel-Lindau protein can adopt only one distribution pattern, which is that two partitions contain zero Q, five partitions contain one Q and one partition contains three Qs (the fourth row in Table 2).

In such a manner, we can quantify each amino acid in wild-type von Hippel-Lindau protein. Thereafter, we can assign these probabilities to each amino acid in the von Hippel-Lindau protein as shown in Figure 1, from which we get the visual sense of how these distribution probabilities go along the von Hippel-Lindau protein, and more importantly we can sum up these distribution probabilities together for all 213 amino acids in the protein.

Actually, the Maxwell-Bolzmann assumption provides us a way to estimate the position of amino acid in a protein, because there is a standard method for the computation using Maxwell-Bolzmann assumption, which saves us from inventing new computational methods. Moreover, the primary structure is the base for higher-level structure, thus any mutation in primary structure would lead to the change in distribution probability, in higher-level structure, and finally the biological function. This is the biological meaning of use of Maxwell-Bolzmann assumption for quantify-

<table>
<thead>
<tr>
<th>Partition 1</th>
<th>Partition 2</th>
<th>Partition 3</th>
<th>Partition 4</th>
<th>Partition 5</th>
<th>Partition 6</th>
<th>Partition 7</th>
<th>Partition 8</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.002403</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.0673</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.0673</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>5.6076e-3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>5.6076e-4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>2.6703e-5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3.3466e-5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0.0421</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>3.3466e-3</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>9.3466e-5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.2523</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.0421</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>0.0280</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5.6076e-3</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>1.8692e-4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1.8692e-4</td>
</tr>
</tbody>
</table>
cational sequence.

In this context, any clinical manifestations related to mutation in proteins would have different distribution probabilities determined by Maxwell-Bolzmann assumption. This is the association between them.

2.4. Quantification of Mutated von Hippel-Lindau Proteins

The calculation in the above subsection is referred to the amino-acid distribution probability before mutation, say, the amino-acid distribution probability in wild-type von Hippel-Lindau protein. Obviously any point mutation leads an amino acid to change to another one, which certainly would change the distribution pattern of both original and mutated amino acids, thus the amino-acid distribution probability would differ for both original and mutated amino acids between before and after mutation.

For example, the missense mutations at the CpG mutation hotspot at codon 167 can mutate arginine “R” to glycine “G”, or glutamine “Q” or tryptophan “W” [13, 46] leading to type I-II, type II and type II von Hippel-Lindau disease, respectively. In above subsection, we have calculated the distribution probability of Qs (Table 2) before mutation, and now we show the calculation of distribution probability after R167Q mutation.

After this mutation, there are nine Qs in the von Hippel-Lindau mutant (Table 3), for which we have

\[
\frac{9!}{5! \times 1! \times 2! \times 0! \times 0! \times 0! \times 0! \times 0!} \times \frac{0! \times 0! \times 0! \times 2! \times 0! \times 1! \times 3! \times 0! \times 3!}{9!} = 0.0197
\]

while its distribution probability before this mutation is 0.0673, so the mutation decreases the distribution probability of Q. On the other hand, there are 20 and 19 Rs before and after this mutation. Their distribution probabilities are 0.0067 and 0.0030 before and after mutation, so this mutation decreases the distribution probability of R, too. The overall effect for this mutation is (0.0030 - 0.0067) + (0.0197 - 0.0673) = -0.0513, that is, the mutation reduces the distribution probability for von Hippel-Lindau protein.

Since von Hippel-Lindau protein functions as whole, we can calculate the change led by the mutation in following way. The sum of all the distribution probability is 19.6114 in wide-type von Hippel-Lindau protein (Figure 1), while the above calculated mutation leads the sum of all the distribution probability to be 19.1731, thus this mutation results in 2.23% decrease in the measure [(19.1731 - 19.6114)/19.6114%].

In this way, we have the quantitative measure for the changed primary structure of von Hippel-Lindau mutants and we also have documented clinical manifestations induced by the mutations of von Hippel-Lindau protein, thus we can build a quantitative relationship between changed structure and clinical outcome.

2.5. Descriptively Probabilistic Relationship

For building quantitative relationship between mutation and clinical outcome, we use the descriptively probabilistic method, as our quantification is the amino-acid distribution probability and each individual mutation related to its clinical outcome is presented as frequency. Therefore, we use the cross-impact analysis to couple
them [35, 47, 48, 49, 50, 51, 52, 53], because the amino-acid distribution probability either increases or decreases after mutation, which is a 2-possibility event, and the clinical outcome either occurs or does not occur after mutation, which is a yes-and-no event. Thereafter, we can use the Bayesian equation to calculate the probability of occurrence of clinical outcome under a mutation.

2.6. Classification of Clinical Outcomes

It is extremely challenging how to use a mathematical modeling to distinguish the clinical outcomes with respect to mutant von Hippel-Lindau protein because of the variety of clinical outcomes. In an effort towards solving this problem, we employ our second quantification, amino-acid pair predictability, whose relational and applications have been published intensively (for reviews, see [2, 3, 4]).

This quantification is based on permutation, and can be calculated in the following way. For example, there are 30 glutamic acids “E” and 20 Rs in von Hippel-Lindau protein, the predicted frequency of amino-acid pair ER would be 3 (30/213×20/212×212=2.817), while we do find three ERs in the protein, so the amino-acid pair ER is predictable. Still, the predicted frequency of EE would be 4 (30/213×29/212×212=4.085), but actually the EE appears nine times in reality. This is the case that the actual frequency is larger than its predicted one. In this manner, we can quantify a protein sequence according to the percentage of how many amino-acid pairs are predictable among all the amino-acid pairs in given protein as well as its mutants. For instance, the predictable portion of amino-acid pairs is 27.54% in wild-type von Hippel-Lindau protein and 31.88% in its P25L mutant.

2.7. Statistics

The data are presented as mean±SD for normal distribution or median with interquartile range for non-normal distribution. The Kruskal-Wallis one-way ANOVA and Chi-square are used for statistical inference, and P < 0.05 is considered significant.

3. RESULTS AND DISCUSSION

After computing amino-acid distribution probability in wild-type von Hippel-Lindau protein and in its 132 mutants, we have 132 changed amino-acid distribution probabilities. Firstly, we can use the cross-impact analysis to build a quantitative relationship between the increase/decrease of distribution probability after mutations and the clinical diagnosis, because the cross-impact analysis is particularly suited for two relevant events coupled together [35, 47, 48, 49, 50, 51, 52, 53].

Figure 2 displays the cross-impact analysis on the relationship between changed primary structure and von Hippel-Lindau disease. At the level of amino-acid distribution probability, P(2) and P(2') are the decreased and increased probabilities induced by mutations, and 53 and 79 mutations result in the distribution probability decreased and increased, respectively. At the level of clinical diagnosis: 1) P(1|2) is the impact probability (conditional probability) that the von Hippel-Lindau disease is diagnosed under the condition of increased distribution probability, and 70 mutations have such an effect. 2) P(T|2) is the impact probability that other disease is diagnosed under the condition of increased distribution probability, and 9 mutations work in such a manner. 3) P(1|2) is the impact probability that the von Hippel-Lindau disease is diagnosed under the condition of decreased distribution probability, and 44 mutations play such a role. 4) P(T|2) is the impact probability that other disease is diagnosed under the condition of decreased distribution probability, and 9 mutations fall into this category. At the level of combined events, we can see the combined results of changed structure and von Hippel-Lindau disease.

Table 4 lists the calculated probabilities with respect to Figure 2, from which several interesting points can be drawn. 1) As P(2') is larger than P(2), a mutation has a larger chance of increasing the distribution probability in von Hippel-Lindau mutant. 2) As P(1|2) is much larger than P(T|2'), a mutation that increases the distribution probability has about nine tenth chance of being von Hippel-Lindau disease. 3) As P(1|2) is much larger than P(T|2'), a mutation that decreases the distribution probability has much larger chance of being von Hippel-Lindau disease.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before mutation</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>After mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Partition</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before mutation</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>After mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

<p>| Table 4. Distribution pattern of glutamines before and after mutation at position 167 in von Hippel-Lindau protein. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>Partition</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before mutation</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>After mutation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

| Table 4. Computed probabilities in reference to the cross-impact analysis in Figure 2. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P(2)=53/132=0.4015 |
| P(2')=1-P(2)=1-0.4015=0.5985=79/132 |
| P(1|2)=70/79=0.8861 |
| P(T|2)=1-P(1|2)=1-0.8861=0.1139=9/79 |
| P(1|2)=44/53=0.8302 |
| P(T|2)=1-P(1|2)=1-0.8302=0.1698=9/53 |
| P(2)=P(1|2)×P(2)=70/79×79/132=0.5363=70/132 |
| P(T)=P(1|2)×P(2)=44/53×53/132=0.3333=44/132 |
| P(T)=P(1|2)×P(2)=9/53×53/132=0.0682=9/132 |
| P(T)=P(1|2)×P(2)=9/53×53/132=0.0682=9/132 |
Secondly, we use the Bayes’ law $P(1|2)=P(2|1)\frac{P(1)}{P(2)}$, which indicates the probabilities of occurrences of two events [54], to determine the probability, $P(1)$, von Hippel-Lindau disease under a mutation, because $P(2)$ and $P(1|2)$ have already been defined in cross-impact analysis, while $P(2|1)$ is the probability that the distribution probability decreases under the condition that the von Hippel-Lindau disease is diagnosed.

As $P(1|2)=44/53=0.8302$ (Table 4), and $P(2|1)=44/(44+70)=0.3860$, 
$P(1)=\frac{P(1|2)P(2)}{P(2|1)}=\frac{0.8302\times 0.4015}{0.3860}=0.8635$, namely, the patient has nine tenth chance of being von Hippel-Lindau disease when a new mutation is found in von Hippel-Lindau protein.

Among patients with von Hippel-Lindau disease, about 40% of mutations are genomic deletions and the rest are predominantly truncating or missense mutations, which do not occur within the first 53 amino acids [55,56]. In this study, we focus on the mutations of von Hippel-Lindau protein. From a probabilistic viewpoint, our results indicate the chance of being diagnosed as the von Hippel-Lindau disease when a new von Hippel-Lindau mutant occurs.

The von Hippel-Lindau disease is characterized by marked phenotypic variability [57,58], due to mosaicism [59], modifier effects [60], and mainly allelic heterogeneity [61]. All these result in complicated clinical classifications. Thus, we use the predictable portion of amino-acid pairs to model the classifications.

**Figure 3** illustrates the classification with respect to the predictable portion of amino-acid pairs. Although there are large overlaps among classifications, our quantification already distinguishes them to some degree. For example, in comparison with von Hippel-Lindau disease, our quantification shows relatively lower in pheochromocytoma and higher in other disorders ($P=0.079$, Kruskal-Wallis one-way ANOVA). The lack of statistical significance is certainly, in part, due to few cases in some groups, however the trend is clear, which paves the way for further classification using more sophisticated mathematical models.

Genotype-phenotype relationships have revealed that a certain number of missense mutations are associated with a high risk of pheochromocytoma but the mutations that totally loss their functions are associated with a low risk. Most patients with type II von Hippel-Lindau disease have missense mutations whereas the large dele-
Figure 3. Predictable portion of amino-acid pairs induced by mutations of von Hippel-Lindau protein in pheochromocytoma (Pheo), von Hippel-Lindau disease and other disorders. The data are presented as median with an interquartile range ($P = 0.079$, Kruskal-Wallis one-way ANOVA).

Figure 4. Distribution of changed amino-acid distribution probability in endocrine and nonendocrine neoplasia induced by mutations of von Hippel-Lindau protein ($P = 0.094$, Chi-square).
tions and truncating mutations predominate in type I families [11,19,62,63]. Many missense mutations causing a type I phenotype are involved in the core hydrophobic residues and were predicted to disrupt protein structure, whereas type II phenotype missense mutations are involved in substitutions at a surface amino acid that does not cause a total loss of function [64,65].

Figure 4 displays the distribution of changed amino-acid distribution probability in endocrine neoplasia (pheochromocytoma, type II von Hippel-Lindau disease) and nonendocrine neoplasia (type I von Hippel-Lindau disease). As can be seen, the mutations that led to the endocrine neoplasia have the trend to increase the amino-acid distribution probability (upper panel), whereas the mutations that led to nonendocrine neoplasia have the effect to either increase or decrease the amino-acid distribution probability (lower panel). The difference between two panels is mainly considered from view of symmetry. As the x-axis is related to the number of von Hippel-Lindau mutations, this figure would be different when more mutations would be found in future, which might provide the much clearer pattern, although we did not find the statistical difference between two panels ($P=0.094$, Chi-square) now.

From a theoretical viewpoint, one could consider to calculate the distribution probability of all 19 potential types of mutations at each position of von Hippel-Lindau protein, and then find the link between mutations and clinical outcomes. However, the amount of computation is huge because it would be equal to $2.369 \times 10^{32}$ mutations ($19^{213}$), which is not only beyond the capacity of any computers, but also beyond the capacity for comparison. Actually, we really know that each position does not have 19 types of potential mutations, because this mutation process is governed by the translation probability between RNA codon and mutated amino acids [66, 67,68]. On the other hand, our study is focused on the documented data rather than the simulated data.

In this study, we use a single value, the sum of all distribution probability to represent the normal von Hippel-Lindau protein and its mutated proteins, respectively, because there is no other way to use a single value dynamically to represent a protein, namely, the value is different when a protein is different. Without such a measure, we cannot model a protein dynamically with its mutations. To the best of knowledge, currently it is only the accession number that can represent a protein uniquely, however it has nothing to do with the protein itself, i.e. composition, length, function, etc.

In general, one would hope to verify this type of study against the real-life cases, which is possible in future although it would deal with a large-scale collaboration because this type of diseases is not frequently seen in clinical settings, for example, the von Hippel-Lindau disease has a birth incidence of about 1 in 36 000 [11,12].

It will take years to verify what a theoretical study finds with fast-speed computational technique. Even, we cannot verify all the theoretical studies, for example, we cannot create another earth without global warming.

The implications of this study include two aspects. 1) The relationship between changed primary structure and changed function is very meaningful, because it provides the dynamic rather than static relationship between mutant protein and its function. This can furthermore provide us the basis for building a dynamic model to predict the new function in mutant proteins. Nevertheless, we need to quantify the proteins in order to build a dynamic model and this study is doing in such a way. 2) From the clinical viewpoint, the classification of von Hippel-Lindau disease as well as many mutation related diseases needs a considerable amount of clinical assays. Our approach can provide a probabilistic estimate for disease classification after determining which amino acid has mutated, because the primary structure of protein is the base for its high-level structure and function.

4. ACKNOWLEDGEMENTS

This study was partly supported by Guangxi Science Foundation (No. 0537012-G and 0991080), and Guangxi Academy of Sciences (project No. 09YJ17SW07).

REFERENCES


New blind estimation method of evoked potentials based on minimum dispersion criterion and fractional lower order statistics

Daifeng Zha

ABSTRACT

Evoked potentials (EPs) have been widely used to quantify neurological system properties. Traditional EP analysis methods are developed under the condition that the background noises in EP are Gaussian distributed. Alpha stable distribution, a generalization of Gaussian, is better for modeling impulsive noises than Gaussian distribution in biomedical signal processing. Conventional blind separation and estimation method of evoked potentials is based on second order statistics or high order Statistics. Conventional blind separation and estimation method of evoked potentials is based on second order statistics (SOS). In this paper, we propose a new algorithm based on minimum dispersion criterion and fractional lower order statistics. The simulation experiments show that the proposed new algorithm is more robust than the conventional algorithm.

Keywords: Evoked potentials (EPs), Alpha stable distribution, Blind source separation, Minimum dispersion (MD), Fractional lower order statistics (FLOS)

1. INTRODUCTION

The brain evoked potentials (EPs) are electrical responses of the central nervous system to sensory stimuli applied in a controlled manner. The EPs have a number of clinical applications including critical care, operating room monitoring and the diagnosis of a variety of neurological disorders [1, 2]. The analysis of EP characteristics is of special interest in many clinical applications, such as the diagnosis of possible brain injury and disorders in the CNS [11, 12]. Thus, the goal in the analysis of EPs is currently the estimation from the several potentials, or even from a single potential. In recent years, signal processing techniques including adaptive filtering, three-order correlation, and singular value decomposition (SVD) have been used in fast estimation of EPs. Independent component analysis (ICA) appeared as a promising technique in signal processing. Its main applications re feature extraction, blind source separation, biomedical signal processing. ICA is based on the following principles. Assume that the original (or source) signals have been linearly mixed, and that these mixed signals are available. Conventional ICA is optimal in approximating the input data in the mean-square error sense, describing some second order characteristics of the data. Nonlinear ICA [3] method related to higher order statistical techniques is a useful extension of standard ICA. The data are represented in an orthogonal basis determined merely by the second-order statistics (covariance) of the input data [4]. Recent studies [5, 6] show that alpha stable distributions is better for modeling impulsive noise, including underwater acoustic, low-frequency atmospheric, and impulsive EEG,ECG, than Gaussian distribution in signal processing. In general, EP signals are always accompanied by ongoing electroencephalogram (EEG) signals which are considered noises in EP analysis. Often the EEG signals are assumed to be Gaussian distributed white noise for mathematical convenience. However, the EEG signals are found to be non-Gaussian in other studies (e.g., [9, 10]). Consequently, EP analysis algorithms developed under the Gaussian EEG assumption may fail or may not perform optimally. Developing EP analysis algorithms without the Gaussian distribution assumption for the background noise thus becomes a key to ensuring the reliability of the analysis results. There are two kinds of noises in the EP signals obtained. The first one is the background EEG noise found in all EP recordings. The second one is the noise introduced by the impact acceleration experiment. An analysis shows that the alpha stable model fits the noises found in the impact acceleration experiment under study better than the Gaussian model [8].

The kind of alpha stable distribution process has no its second order or higher order statistics. It has no close form probability density function so that we can only describe it by its characteristic function:

$$\varphi(t) = \exp \left[ j \mu t - \gamma |t|^\alpha \right] + j \beta \text{sgn}(t) \varphi(t, \alpha)$$  (1)

Where $\varphi(t, \alpha) = \frac{\tan(\pi \alpha / 2)}{2} \log |t|$ if $\alpha = 1$,

$$\varphi(t) = \exp \left[ j \mu t - \gamma |t|^\alpha \right] + j \beta \text{sgn}(t) \varphi(t, \alpha)$$  (1)

Where $\varphi(t, \alpha) = \frac{\tan(\pi \alpha / 2)}{2}$ if $\alpha = 1$,
\( -\infty < \xi < \infty , \delta > 0, 0 < \alpha \leq 2, -1 \leq \beta \leq 1. \) The characteristic exponent \( \alpha \) determines the shape of the distribution. Especially, if \( \alpha = 2 \), it is a Gaussian distribution. The dispersion \( \delta \) plays a role analogous to the variance of the second order process. \( \beta \) is the symmetry parameter and \( \mu \) is the location parameter. The distinct characteristics of lower order stable process are its impulsive waveform and the thick tail in its distribution function. Due to the thick tails, lower order stable processes do not have finite second or higher-order moments. This feature may lead all second order moment based algorithms to fail or to function sub-optimally. The typical Evoked potentials are shown in Figure 1.

**2. DATA MODEL**

In the following, we present the basic data model used in both PCA and the source separation problem plotted in Figure 2, and discuss the necessary assumptions. We assume that \( P \) signals \( s(n,i) = 1,2,...P \) are non-coherent, statistically independent. The noiseless linear ICA model with instantaneous mixing may be described by the equation

\[
x(n) = As(n) = \sum_{i=1}^{P} a_i s_i(n)
\]

(2)

where \( x(n)=[x_1(n),x_2(n),...,x_M(n)]^T \) are observed signals, \( S(n)=[s_1(n), s_2(n),...,s_M(n)]^T \) are the source signals containing alpha stable distribution signals or noise which are supposed to be stationary and independent, and \( A \) is an unknown mixing matrix. Our goal is to estimate \( S \) from \( X \), with appropriate assumptions on the statistical properties of the source distributions. The solution is

\[
Y(n) = WZ(n)
\]

(3)

where \( W \) is the de-mixing matrix, \( Z(n) \) is the whitening vector. The general ICA problem requires \( A \) to be an \( N \times P \) matrix of full column rank, with \( M \geq P \), i.e., there are at least as many mixtures as the number of independent sources. In this paper, we assume an equal number of sources and sensors to make calculation simple. We can write the signal model in matrix form as \( X = AS \). Here \( X \) is observation data matrix, \( S \) is source signals data matrix, mixture matrix \( A \) is unknown.

**3. WHITENING BY NORMALIZED COVARIANCE MATRIX**

Generally, it is impossible to separate the possible noise in the input data from the source signals. In practice, noise smears the results in all the separation algorithms. If the amount of noise is considerable, the separation results are often fairly poor. Some of the noise can usually be filtered out using standard PCA if the number of mixtures is larger than the number of sources [13].

We introduce here a two-step separation method that achieves the BSS through minimization of a dispersion criterion. The first step is a whitening procedure that orthogonalizes the mixture matrix. Here we search for a matrix \( B \) which transforms mixing matrix \( A \) into a unitary matrix. Classically, for a finite variance signal, the whitening matrix is computed as the inverse square root of the signal covariance matrix. In our case, impulsive EEG noises have infinite variances. However, we can take advantage from the normalized covariance matrix.

Theorem 1 [7]: Let \( X=[x(1),x(2),...,x(N)] \) be a stable process vectors data matrix, then normalized covariance matrix of \( X \)

\[
\Gamma_x = \frac{X X^T}{N \cdot \text{Trace}(X X^T / N)}
\]

(4)

Converges asymptotically to the finite matrix when \( N \rightarrow \infty \), i.e.,

\[
\lim_{N \rightarrow \infty} \Gamma_x = ADA^T, D = \text{diag}(d_1,d_2,...d_M)
\]

where \( d_i = \lim_{N \rightarrow \infty} \Delta_j \sum_{j=1}^{P} \Delta_j \|a_j\|^2, \ a_j \) is column of \( A, \Delta_j = \sum_{n=1}^{N} x_j^2(n) / N \).

Theorem 2: We have eigen-decomposition of \( \Gamma_x \) as \( \Gamma_x = U \Omega^2 U^T \) and we can obtain whitening matrix \( B = U \Omega^2 U^T \), then \( Z = BX \) is orthogonal.

![Figure 1. Evoked potentials (Left: without noises; Right: with noises).](image-url)
which are suitable for alpha stable noise environments in the following, we discuss and propose separation algorithms. ral blind separation algorithms have been proposed. In the separating matrix, we maximize the output variances.

Let objective function be }

\[
\text{(5)}
\]

Here the Lagrange multiplier is } \lambda_j \text{. These can be taken into account by multiplying (6) by } W_j^T \text{ from left. We can obtain }

\[
\lambda_j = -W_j^T E[Z(n) | Z^T(n) W_i | p=2 \text{ cond}(Z^T(n) W_i)].
\]

Inserting these into (6), we can get

\[
\hat{V}J(W_i) = [1 - \sum_{j=1}^{p} W_j W_j^T] E[Z(n) | Z^T(n) W_i | p=2 \text{ cond}(Z^T(n) W_i)].
\]

A practical gradient algorithm for optimization problem (5) is now obtained by inserting (7) into

\[
\text{W}_i(n+1) = \text{W}_i(n) - \mu(n) \hat{V}J(W_i(n)),
\]

where } \mu(n) \text{ is the gain parameter. The final algorithm is thus }

\[
\text{W}_i(n+1) = \text{W}_i(n) - \mu(n) \left[ \frac{1}{N} \sum_{n=1}^{p} \text{W}_i(n) \text{W}_j^T(n) [Z(n)] - Z^T(n) \text{W}_i(n) \right] \text{p=2 \text{ cond}(Z^T(n) \text{W}_i(n))}
\]

As } y_i(n) = Z^T(n) \text{W}_i(n), (8) \text{ can be written as follow }

\[
\text{W}_i(n+1) = \text{W}_i(n) - \mu(n) \left[ \frac{1}{N} \sum_{j=1}^{p} y_j(n) \text{W}_j(n) \right] \text{p=2 \text{ cond}(y_i(n))}
\]

Let } g(t) = t \text{p=2 \text{ cond}(t)}, \text{ then } g(t) \text{ is appropriate PCA network nonlinear transform function for lower order alpha stable distribution impulse noises.}

Considering that during the iteration error item of gradient } 1 - \sum_{j=1}^{p} \text{W}_j(n) \text{W}_j^T(n) \text{ might be zero instantaneously, we modify (9) in order to improve robustness of algorithm as }

\[
\text{W}_i(n+1) = \text{W}_i(n) - \mu(n) \left[ \frac{1}{N} \sum_{n=1}^{p} y_i(n) \text{W}_i(n) - \mu(n) \left[ \frac{1}{N} \sum_{j=1}^{p} y_j(n) \text{W}_j(n) \right] \right] \text{p=2 \text{ cond}(y_i(n))}
\]

Thus } \text{W}_1, \text{W}_2, ..., \text{W}_p \text{ can be obtained. Let } Y(n) = [y_1(n), y_2(n), ..., y_p(n)]^T, W = [W_1, W_2, ..., W_p]. \text{ For whole network, solution to } W \text{ and optimization problem is }

\[
\text{W}_i(n+1) = \text{W}_i(n) - \mu(n) \left[ \text{Z}(n) - \text{W}(n) g(Y(n)) \right] g(Y^T(n))
\]

5. Performances Analysis

Different nonlinear function can be applied to different blind signal separation problem. Many popular functions are } g(t) = \text{sign}(t) \text{ and } g(t) = \text{tanh}(t) \text{ corresponding to thodelinear exponential distribution } \frac{1}{2} \exp(-|x|) \text{ and the inverse-cosine-hyperbolic distribution } \frac{1}{\pi} \frac{1}{\cosh(x)}, \text{ respectively. For the class of symmetric normal inverse distribution }
Gaussian (NIG), it is straightforward to obtain according to [14] \( g(t) = -\alpha t \left( K_2(\alpha \sqrt{\delta^2 + t^2}) \right) \), where \( K(.) \) and \( K_2(.) \) are the modified Bessel function of the second kind with index 1 and 2.

For lower order alpha stable distribution noise has no second order or higher order moment, we must select appropriate nonlinear function \( g(t) = |t|^{p-2} \text{conj}(t) \) (\( p < \alpha \)). If \( t \) is real data, then \( g(t) = |t|^{p-1} \text{sign}(t) \). If \( p=1 \), then \( g(t) = \text{sign}(t) \).

Figure 3 shows the nonlinear function of alpha stable distribution for different \( \alpha \).

We start from the learning rule (12), and we assume that there exists a square separating matrix \( H^T \) such that \( U(n) = H^T Z(n) \). The separating matrix \( H^T \) must be orthogonal. To make the analysis easier, we multiply both sides of the learning rule (12) by \( H^T \). We obtain

\[
H^T W(n+1) = H^T W(n) + \mu(n) [H^T Z(n) - H^T W(n) g(W(n)^T Z(n))] g(Z(n)^T W(n))
\]

For the sake of \( H H^T = I_p \), we can get

\[
H^T W(n+1) = H^T W(n) + \mu(n) [H^T Z(n) - H^T W(n) g(W(n)^T H H^T Z(n))] g(Z(n)^T H H^T W(n))
\]

Define \( Q(n) = H^T W(n) \), \( W(n) = (H^T)^{+} Q(n) \), (14) is written as

\[
Q(n+1) = Q(n) + \mu(n) [U(n) - Q(n)] g(Q(n)^T U(n))] g(Q(n) U(n)^T)
\]

Geometrically the transformation multiplying by the orthogonal matrix \( H^T \) simply means a rotation to a new set of coordinates such that the elements of the input vector expressed in these coordinates are statistically independent.

Analogous differential equation of (15) is obtained as matrix form:

\[
dQ/dt = E[Qg(U^T Q)] - E[g(Q^T U)g(U^T Q)]
\]

According to [15], we can easily prove that (16) has stable solution. For the sake of \( Q=H^T W \), thus \( W = (H^T)^{+} Q \) is asymptotic stable solution of (12). Figure 4 shows the stability and convergence of algorithm based on SOS and FLOS. From Figure 4, we know the algorithm based on FLOS has better stability and convergence than the algorithm based on SOS.

6. EXPERIMENTAL RESULTS

From Section I we know that the noise for EP could be a lower order stable process. Through computer simulations, we will demonstrate the effectiveness of the proposed algorithm under alpha stable noise conditions. We use correlation coefficient as follows to evaluate the performances of the proposed algorithms:

\[
\tau(s_j, y_j) = \frac{\sum_{n=1}^{N} s_j(n)y_j(n)}{\sqrt{\sum_{n=1}^{N} s_j^2(n)\sum_{n=1}^{N} y_j^2(n)}}
\]

**Experiment 1**

Two independent sources are linearly mixed. One is the periodical noise free EP signal, and the period is 128 points, the sampling frequency is 1000Hz. The other is an alpha stable non-Gaussian noise with \( \alpha = 1.7 \). Two algorithms are used in the experiment, including: (1) SOS with nonlinear function \( g(t) = \text{tanh}(t) \); (2) FLOS with \( g(t) = t|t|^{p-2} \text{conj}(t) \), respectively. Figure 4 shows the stability and convergence of algorithm based on SOS and FLOS. We know the algorithm based on FLOS has better stability and convergence than the algorithm based on SOS.

We can get signals waveforms in time domain shown in Figure 5, where (a) and (b) are source signals, (c) and (d) are separated signals based on SOS, (e) and (f) are separated signals based on FLOS. For FLOS algorithm,
Figure 4. The stability and convergence.

Figure 5. Separating results: (a)-(b) are the source signals. (c)-(d) are the separated signals with SOS. (e)-(f) are the separated signals with FLOS.

Figure 6. Separating results: (a)-(b) are the source signals. (c)-(d) are the separated signals with SOS. (e)-(f) are the separated signals with FLOS.
Table 1. Comparison between the two algorithms.

<table>
<thead>
<tr>
<th>Iteration times</th>
<th>Correlation coefficient (FLOS)</th>
<th>Correlation coefficient (SOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EP noise</td>
<td>EP</td>
</tr>
<tr>
<td>50</td>
<td>0.1244</td>
<td>0.1044</td>
</tr>
<tr>
<td>100</td>
<td>-0.3450</td>
<td>-0.3050</td>
</tr>
<tr>
<td>150</td>
<td>0.4378</td>
<td>0.4378</td>
</tr>
<tr>
<td>200</td>
<td>0.6766</td>
<td>0.7706</td>
</tr>
<tr>
<td>250</td>
<td>-0.9291</td>
<td>-0.9091</td>
</tr>
<tr>
<td>300</td>
<td>-0.9287</td>
<td>-0.9107</td>
</tr>
<tr>
<td>350</td>
<td>-0.9293</td>
<td>-0.9113</td>
</tr>
<tr>
<td>400</td>
<td>0.9295</td>
<td>0.9195</td>
</tr>
<tr>
<td>450</td>
<td>0.9299</td>
<td>0.9292</td>
</tr>
<tr>
<td>500</td>
<td>-0.9501</td>
<td>-0.9593</td>
</tr>
</tbody>
</table>

Figure 7. The correlation coefficients of EP and noise.

the correlation coefficient between the separated and source EP signals is 0.9213, and the correlation coefficient between the separated and source alpha stable non-Gaussian noises is –0.9098.

Experiment 2

We repeat simulations when GSNR is 20dB. Two independent sources are linearly mixed. One is the periodical noise free the brain evoked potential (EP) signal, and the period is 128 points, the sampling frequency is 1000Hz. The other is an alpha stable non-Gaussian noise with $\alpha = 1.7$. Two algorithms are used in the experiment, including: (1) SOS with nonlinear function $g(t) = \tanh(t)$; (2) FLOS with $g(t) = t|t|^{-2} \text{conj}(t)$, respectively. We can get signals in time domain shown in Figure 6, where (a) and (b) are source signals, (c) and (d) are separated signals based on SOS, (e) and (f) are separated signals based on FLOS. For FLOS algorithm, the correlation coefficient between the separated and source EP signals is –0.9213, and the correlation coefficient between the separated and source alpha stable non-Gaussian noises is –0.9098.

Experiment 3

Separate the mixed signals again with the new FLOS algorithm and conventional SOS algorithm, respectively. And the results of 10 independent experiments are shown in Table 1. The correlation coefficients of EP and of the noise are calculated at some iteration times and plotted in Figure 7. From Table 1, we get that the performance of the new algorithm is better than the Conventional algorithm.

7. CONCLUSION

Alpha stable distributions, is better for modeling impulsive noise than Gaussian distribution in biomedical signal processing. Conventional blind separation and estimation method of evoked potentials is based on second order statistics. In this paper, we modify conventional algorithms and analyze the stability and convergence performance s of the new algorithm. From above simulation, we can easily obtain the following conclusions: the proposed class of algorithm of estimation of evoked potentials based on FLOS is more robust than conventional algorithms based on SOS so that its separation capability is greatly improved under both Gaussian and fractional lower order stable distribution noise environments.

ACKNOLEDGEMENT

This work is supported by the National Science Foundation of China under Grant 60772037 and the Science Foundation of Department of Health of Jiangxi province under Grant 20072048.

REFERENCES


Visualization of protein structure relationships using constrained twin kernel embedding

Yi Guo, Jun-Bin Gao, Paul Wing Hing Kwan & Kevin Xinsheng Hou

ABSTRACT
In this paper, a recently proposed dimensionality reduction method called Twin Kernel Embedding (TKE) [10] is applied in 2-dimensional visualization of protein structure relationships. By matching the similarity measures of the input and the embedding spaces expressed by their respective kernels, TKE ensures that both local and global proximity information are preserved simultaneously. Experiments conducted on a subset of the Structural Classification Of Protein (SCOP) database confirmed the effectiveness of TKE in preserving the original relationships among protein structures in the lower dimensional embedding according to their similarities. This result is expected to benefit subsequent analyses of protein structures and their functions.

1. INTRODUCTION
Recent years have seen phenomenal advances in dimensionality reduction (DR) methods that are widely applied in bioinformatics [26, 14, 6, 16], biometrics [19, 4, 15, 17], robotics [11], etc. The rapid growth of DR methods stems from the need to reduce the complexity of the problem at hand. The target of these methods is mainly to find the corresponding counterparts of the input data in a much lower dimensional space without incurring significant information loss. The low dimensional representation can be used in subsequent procedures such as classification, pattern recognition, and so on. For example, a medium sized protein structure typically has a few thousands of degrees of freedom which is naturally residing in very high dimensional space. It causes the so-called “curse of dimensionality” problem which will not only drastically increase the computational complexity of the learning algorithms but also require large storage space, leading to very slow indexing and searching speed in a large scale database in the sequel. Through DR, most of the redundant dimensions can be removed and the degrees of freedom left are then input to the subsequent discriminant and classification tasks with considerable simplification.

Another advantage of using DR is the 2- or 3-dimensional mappings of the original data can be visualized in an Euclidean space that can facilitate interpreting the relationships among data by the researchers. Normally, the relationships among a set of protein structures are typically represented in the form of trees derived by hierarchical clustering. However, this representation only provides some hints on the evolutionary distances between protein structures. This limitation motivates our applying to the application of DR methods in visualizing the similarity relationships among protein structures.

In this paper, we will propose a new DR algorithm called Constrained Twin Kernel Embedding (CTKE) based on Twin Kernel Embedding (TKE) [10] to achieve the above target. To provide the necessary background knowledge, we will first give a brief review of DR methods on protein structures in the next section, followed by an introduction of TKE and related topics involved in this new algorithm. Then the CTKE algorithm will be introduced, which integrates the similarity matching by TKE with the objective function used in LS-SVM [23]. Experiments on real protein structure data will be presented and finally we summarize this paper with a conclusion.

2. THE RELATED WORKS
DR methods can be categorized into Linear DR methods (LDR) such as Principal Component Analysis (PCA) [13], Linear Discriminant Analysis (LDA) [7] and NonLinear DR methods (NLDR) such as ISOMAP [25], Laplacian Eigenmaps (LE) [3], Locally Linear Embedding (LLE) [20], etc. LDR methods have been widely used in bioinformatics due to their simplicity. For example, Teodoro et al. [27] applied PCA to transform the original high dimensional protein motion data into a lower dimensional representation that captures the dominant modes of motions of the protein. However, the linearity assumption on which linear methods are constructed does not hold in most cases. In [5], Das et al.
have to be vectorial. Any structured data like protein structures can be properly processed by specially designed kernel functions. This advantage avoids the information loss during vectorization. TKE is constructed on the basis of this kind of similarity metric. The objective function is totally different from Sammon’s mapping and furthermore the dis/similarity metric for images is not limited to simple Euclidean distance, but a kernel function to capture the nonlinearity.

3. TWIN KERNEL EMBEDDING

Without loss of generality, the following notations will be adopted. The data in the input space $y$ are denoted by $y_i$ ($i = 1, \ldots, N$) while $x_i$ ($i = 1, \ldots, N$) their embeddings in a low dimensional space or the so-called latent space $X$. Notice that here the $y_i$ will be the protein structures. The term “embeddings” is from the manifold learning literature which means the images of the input data or equivalently the embedded data. In addition, $Y$ and $X$ will be used to denote respectively the set of input objects and the set of embedded objects. If the objects were vectorial, $Y$ (and $X$) would denote a matrix consisting of rows of vectors. Furthermore, $ab$ denotes the inner product of two vectors $a$ and $b$.

The Twin Kernel Embedding (TKE) preserves the similarity structure of input data in the latent space by matching the similarity relations represented by two kernel Gram matrices, i.e. one for the input data and the other for their embeddings by simply minimizing the objective function

$$-\text{Vec}K_y \cdot \text{Vec}K_y,$$

(1)

where Vec is the vec operator on matrix (to stack all the columns of the matrix to make a long vector) and $K_y$ and $K_x$ are the kernel Gram matrices derived from valid Mercer kernel functions $k_y(\cdot, \cdot)$ and $k_x(\cdot, \cdot)$ [21] defined on the input data and embeddings respectively.

The idea is to preserve the similarities among the input data and reproduce them in the lower dimensional latent space expressed again in similarities among embeddings. To make this point clearer, we can simply regard $-\text{Vec}K_y \cdot \text{Vec}K_x$ as a linear kernel (liner kernel is defined as $k(a, b) = a \cdot b$) which is a measure of similarity of the variables involved in the kernel function. The larger the value of the kernel, the more similar these two variables are. As a result, we minimize (1) to make $K_y$ and $K_x$ as similar as possible.

To avoid any trivial solutions to (1), two regularization terms on the kernel and embeddings are introduced and the objective function in (1) becomes

$$L = -\text{tr}(K_y \cdot K_y) + \lambda_1 \text{tr}(K_x \cdot K_x) + \lambda_1 \text{tr}(XX^T)$$

(2)

where we use the fact that $\text{Vec}K_y \cdot \text{Vec}K_x = \text{tr}(K_y \cdot K_x)$. The second term is a ridge regularizer the kernel to make sure that the norm of the kernel is controlled. This

successfully projected the folding free-energy on a few relevant coordinates by using a typical nonlinear method, ISOMAP, to correctly identify the transition-state ensemble of the reaction based on the fact that empirical reaction coordinates routinely used in protein folding studies cannot be reduced to a linear combination of the Cartesian coordinates.

Because of the power of the NLDR methods, they have also been applied to visualizing the relationships between protein structures. Hanke and Reich [12] employed the Kohonen self organizing maps, a special form of neural networks as a visualization tool for the analyses of protein structure similarity by converting the sequences into a characteristic signal matrix. In [1], by using the pairwise similarity index between two sequences, Sammon maps projected the sequences onto a display plane in such a way that the Euclidean distances between the images approximate as closely as possible the corresponding values in the original sequence space. The metric used to measure the similarity between two protein structures was based on the individual residue similarities derived from a series of amino acid exchange matrices. Further, a modified nonlinear Sammon projection was developed in [2] to display the relationships among protein structures based on their amino acid composition.

Recently, a new method called Stochastic Proximity Preserving (SPE) was introduced into this field by Far num et al. in [6]. SPE preserves only the local relationships among closely related sequences to avoid a drawback of those global methods such as MDS that underestimates the proximity of sequences since all pairwise distances are included in the algorithm, leading to erroneous results. To emphasize the proximity, SPE first applies a neighborhood filtering procedure to the similarity matrix (the similarity metric used in SPE is identical to that used in [1]). Then the filtered similarity matrix is input into the Sammon’s nonlinear mapping to obtain the final result.

From the discussion above, we can clearly see that there are three important components in DR: dis/similarity metric for the input data (protein structures in this paper), the objective function (the core of the algorithm) and the dis/similarity metric for images (the corresponding low dimensional representation of the protein structures) which is usually the Euclidean distance. These DR methods are trying to preserve the similarity metric of the protein structures as much as possible and reproduce it in a human interpretable space. The dis/similarity metrics used in those methods mentioned above are based on proteins and their structure-related evaluators while the objective functions are from Sammon’s mapping.

In computational biology, the similarity metric known as kernel functions that are both powerful and promising is gaining much attention. An important advantage of a kernel function is that the form of the input data does not
can avoid solutions that simply let the elements in \( K_x \) go to infinity. The third term imposes a heavy penalty on too large a norm for the embeddings which ensure that their coordinates are relatively small. \( \lambda_i \) and \( \lambda_j \) are tunable parameters to control the strength of the regularization and are assumed to be positive.

In order to capture the nonlinear structure, \( k_x(x, x') \) should be chosen to be nonlinear. Normally, we use the RBF kernel
\[
k_x(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2}\right)
\] (3)
where \( \gamma \) and \( \sigma \) are positive hyperparameters. Nonetheless, other kernels can also be applied here. The selection of RBF kernel is for its connection to the Euclidean distance which has a geometric understanding of the embeddings. There is no closed form solution for \( \mathbf{X} \) and hence an optimization procedure like gradient descent based algorithm for optimization should be employed provided that \( k_x(x, x') \) is differentiable. The initialization of \( \mathbf{X} \) is also required to start the optimization process. KLE [9], KPCA [22] and other methods that can work with kernels could be utilized here for initializing \( \mathbf{X} \).

The dimension of the embeddings which is normally 2 is assigned according to the application. A by-product of this optimization process is that we can get the optimal hyper-parameters (such as \( \gamma \) and \( \sigma \) if RBF kernel is used) of the kernel function \( k_x(x, x') \) as well. It ensures that the kernel we pick up is well adjusted.

TKE is designed to preserve locality and non-locality at the same time. This is done by filtering of the entries in \( K_y \). Not all the entries remain the same in the optimization process but those that convey most of the similarity information of the input data. This filtering process is fulfilled by performing k-nearest neighbor selecting procedure on \( K_y \). Given an object \( y_i \) in the input space, only those objects whose similarities (in the sense of kernel values) to \( y_i \) are in the \( k \) nearest neighbors that are selected to retain their original values while all others are set to 0. The variable \( k \) \((>1)\) in \( k \)-nearest neighboring controls the neighborhood that the algorithm will preserve. It can be interpreted as constructing a weighted adjacency graph which performs in feature space and the weights on edges are evaluated by a kernel as that in KLE. Similar procedure is involved in SPE as well which uses the neighborhood radius to ensure the proximity preserving property. Because TKE tries to match \( K_x \) to \( K_y \) and RBF kernel (3) cannot give a value of 0 except that two points in the latent space are very far apart, TKE seeks a solution that keeps the points in the same neighborhood close while makes the points not in the same neighborhood be very far. However, TKE also works without filtering in which case TKE will preserve all pairwise similarities and will become a global approach simply.

In addition to the fact that TKE outperforms other methods such as KPCA, KLE etc, an elegant feature of TKE is that it uses only the pairwise similarities since in its objective function, only the kernel Gram matrix of the input data is required. Through TKE, any kind of data can be visualized in lower dimensional space as long as an appropriate kernel is defined for them. As a result, a kernel Gram matrix on the input data \( \mathbf{K}_y \) and an initialization for \( \mathbf{X} \) will be adequate for TKE to find the optimal embeddings.

4. CONstrained TWIN Kernel Embedding

It is clear from observing (2) that there is no explicit connections between input data and their corresponding embeddings in TKE except the similarity preserving. As such, the information hidden in the input data is neglected. For example, if the input data actually stay on or near to a smooth manifold embedded in the ambient space, the location of the input data can be explored to predict the coordinates of the embeddings on the manifold. Furthermore, TKE can only find the optimal embeddings for currently presented data as we can see from its objective function. To address these problems, we introduce the constraints reflecting the relationship between input data and embeddings into TKE by following steps. We first define a mapping function \( f : y \rightarrow x \) and then incorporate it into the objective functional of TKE as regularization terms. Finally the optimal embeddings \( \mathbf{X} \) and the \( f \) will be searched via conjugate gradients algorithm.

We can start from the kernel feature mapping directly and incorporate it as the core part of the LS-SVM (the dual form of the objective function of LS-SVM) into TKE as what has been done in [24]. We minimize the following objective function similar to that of LS-SVM with equality constraints
\[
J = \frac{u}{2} \sum_{j=1}^{d} w_j^T w_j + \frac{h}{2} \sum_{j} e_j^2
\] (4)
\[
s.t. \; x_{ij} = \phi_j(y_i) + e_{ij}
\] (5)
corresponding to the maximum margin (the first term) and least square errors (the second term) in (4) where \( u \) and \( h \) are adjustable parameters. \( \phi_j \)’s are the feature mappings which map the input \( y_i \) into Hilbert space where the inner product is defined. \( w_j \) is a column vector having the same dimension as the Hilbert space. We can regard the \( x_i \) in (5) as the projection of \( \phi(y_i) \) onto a subspace parameterized by \( \phi_j \)’s. Because the difference of the constraints, it does not have the same geometrical interpretation of LS-SVM because the original constraints are inequalities reflecting the correct classification hyperplanes while here they are just feature map-
ping which builds the relation between input data $y_i$ and its counterpart $x_i$ in low dimensional space. In (4), we are minimizing the error $e_{ij}$ in reconstruction of $x_i$ essentially however it should be done with $o_{ij}$ properly constrained. This happens to have the form of the LS-SVM objective function. To solve (4) with the equality constraints (5), we can use Lagrange multipliers as

$$L = \frac{u}{2} \sum_{i,j} y_i^T w_j + \frac{h}{2} \sum_{i} y_i^2 + \sum_{i,j} a_{ij} (x_i - w_i^T f_j(y_i) - e_{ij})$$

(6)

From the saddle points, \(\frac{\partial L}{\partial \omega_{ij}} = 0\), \(\frac{\partial L}{\partial e_{ij}} = 0\) we have

$$\frac{\partial L}{\partial \omega_{ij}} = \sum_i a_{ij} \phi (y_i) = 0 \Rightarrow \omega_{ij} = C \sum_i a_{ij} \phi (y_i)$$

$$\frac{\partial L}{\partial e_{ij}} = \eta_{ij} - \omega_{ij} = 0 \Rightarrow e_{ij} = \eta_{ij}$$

Substitute them back into (6) and eliminate $\omega_{ij}$ and $e_{ij}$ we have the dual problem to be maximized according to the min-max duality

$$L = \frac{1}{2} \sum_{i,j} \left( \sum_{i} a_{ij} f_i(y_i) \right)^T \left( \sum_{i} a_{ij} f_i(y_i) \right) + \frac{h}{2} \sum_{i} y_i^2 + \sum_{i,j} \left( \frac{1}{2} a_{ij} \right)^2$$

$$+ \frac{1}{2} \sum_{i} a_{ij} K_{ij} \sum_{i,j} a_{ij} y_i$$

(7)

where $a_j = (a_{1j}, \ldots, a_{2j})^T$, $\phi(y_m)^T \phi(y_i) = k_j(y_m, y_i)$ to which the “kernel trick” applies. $k_j(\cdot, \cdot)$ is the kernel associated with the kernel mapping and $K_j$ is the Gram matrix derived from $k_j(\cdot, \cdot)$ accordingly. We then maximize the above dual problem with respect to $a_{ij}$ instead of $\omega_{ij}$ and $e_{ij}$. From the discussion above we see that the $x_{ij}$’s are free variables. To limit the choice of them, we combine (7) with the objective functional of TKE to incorporate the similarity preserving as

$$L = \sum_{i,j} k_j(y_m, y_i) j_i(x_j, x_j) + \frac{1}{2} \sum_{i,j} k_j(y_i, y_j) + 1 \sum_{i} x_i x_i^T$$

$$+ \frac{1}{2} \sum_{i} a_{ij} K_{ij} + \frac{1}{2h} \sum_{i} \sum_{j} a_{ij}^2 + \sum_{i} a_{ij} y_i$$

(8)

where we turn the maximization of (7) to minimization aligned with the TKE objective functional. Here we see that the terms related to $y_i$ are expressed by kernel $k_j(\cdot, \cdot)$. Therefore, this revised objective function is still non-vectorial data applicable. Again we express (8) into matrix form to facilitate the differentiation and let $k_j(\cdot, \cdot) = k_j(\cdot, \cdot)$ for simplicity

$$L = -Tr[K, K_j] + l_tr[K_j^2] + l_tr[XX^T]$$

$$+ \frac{1}{2u} tr[A^T K, A] + \frac{1}{2h} tr[A^T A] - tr[A^T X]$$

(9)

where

$$A = \begin{bmatrix} a_{ij} & 0 \\ M & M \\ a_{ij} & \frac{1}{h} a_{ij} \end{bmatrix}$$

Hence we minimize $L$ in (8) with respect to $A$, $X$ and kernel hyperparameters of $k_j(\cdot, \cdot)$. If we substitute the saddle point solution back into the equality constraints (5), the following mapping function is handy to predict new input samples

$$x_{ij} = \frac{1}{u} \sum_{m} a_{ij} k_j(y_m, y_i) + \frac{1}{h} a_{ij}$$

(10)

The errors $\frac{1}{\eta} a_{ij}$ can be neglected since the values of the errors are very small compared with $X$ after optimization.

To apply the conjugate gradient algorithm, the derivatives of $L$ with respect to $X$ and $A$ are given by

$$\frac{\partial L}{\partial X} = 2 \lambda K_x - A$$

and,

$$\frac{\partial L}{\partial A} = 2 \lambda K_x - K_y$$

(11)

$X$ is still initialized by KPCA or KLE to obtain pure non-vectorial data applicability. Initial $A$ is from the solution of

$$\frac{\partial L}{\partial A} = 0$$

after $X$ is known. We have

$$A = (\frac{1}{u} K_x + \frac{1}{\eta} I)^{-1} X$$. It implies that we could alternately update $X$ and $A$ in optimization, however we still use conjugate gradient algorithm to update them at the same time. Because the mapping function defined between input space and latent space acts as constraints, we call this algorithm Constrained TKE (CTKE). It is noticeable that in CTKE, we use $K_x$ to construct the mapping function, so we do not have to filter $K_y$ before commencing the algorithm.

5. EXPERIMENTAL RESULTS

Experiments were conducted on the SCOP (Structural Classification Of Protein). This database is available at http://scop.mrc-lmb.cam.ac.uk/scop/. The database comes with a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure. 292 protein structures from different protein superfamilies and families are extracted for the test. The kernels we used are from the family of the so-called alignment kernels whose thorough analyses can be found in [18]. The corresponding kernel Gram matrices are available on the website of the paper as supplements and were used directly in our experiments.

5.1 Parameters Setting

Both CTKE and TKE have parameters to be determined beforehand. Through empirical analyses (performed a set
of experiments on the same data set varying only the parameters, we found these algorithms are not sensitive to the choice of the parameters, so long as the conjugate gradient optimization can be carried out without premature early stop. So we use the following parameters in the $k = 10$ in $k$ nearest neighboring: for CTKE, $\lambda_k = 0.05$, experiments. For TKE, $\lambda_k = 0.005$, $x = 0.001$, $\eta = 0.01$, $\nu = \eta = 0.5$. The minimization will stop after 1000 iterations or when consecutive update of the objective function is less than $10^{-7}$. $k_s(\cdot, \cdot)$ is the RBF kernel and initialization is done by KPCA for both CTKE and TKE.

Figure 1. Visualization results of protein structures.
5.2 Proteins on 2-D Plane

The proteins from the same families are expected to be close in the 2-dimensional space with overlappings indicating similar proteins but actually from different families. The results are plotted in Figure 1. The results of other methods (kernel applicable methods) are also presented for comparison in Figure 2. Each point (denoted as a shape in the figures) represents a protein. The same shapes with the same colors are the proteins from same families while the same shapes with different colors represent the proteins from different families but of the same superfamilies. All the figures in this paper share the same legends as those in Figure 1 (b).

Both TKE and CTKE reveal the fact that proteins from the same families congregate together as clusters while KPCA and KLE fail to reveal it. For example, in Figure 1 (a) and (b), almost all of the proteins from the globin family gathered at the bottom left corner indicating similar structures. Interestingly, CTKE and TKE also reveals the fact that the proteins from the same superfamily but different families are similar in structure, which is reflected in the 2-dimensional plane that the corresponding groups (families) are close if they are in the same superfamily. For instance, note that the proteins from ferritin and ribonucleotide reductase-like families (blue diamonds and violet diamond respectively in the figure), they are close in the group level. SPE has comparable performance visually. But the overlapping in the middle shows that it cannot distinguish some families clearly.

In order to further quantify the results, we use the “purity” [8] to evaluate some of these methods. It uses the fraction of the number of samples from the same class as given point in a neighborhood with size \( n \). The purity is the average of the fraction over all points. The higher the purity, the better the quality of the clusters. Intuitively, this standard provides an objective judgement from the classification point of view and the method with better performance is expected to have a higher purity.

Both TKE and CTKE have high purity, indicating that they are effective in clustering similar proteins. In contrast, KPCA and KLE have lower purity, suggesting that they are less effective in this task. SPE also shows good performance, but there is some overlapping in the middle, indicating that it may not always be able to distinguish between families.

![Figure 2. The result of other methods with MAMMOTH kernel. Like TKE, SPE can work globally and locally depending on whether the kernel Gram matrix is filtered first by k-nearest neighboring. SPE global worked with the whole kernel Gram matrix and SPE local filtered \( K_p \) with 100-nearest neighboring in this experiment](image_url)
purity has the potential to achieve better classification rate. It is noticeable that for SPE and KLE with parameter $k$, we did multiple experiments to choose the optimal value of $k$ corresponding to the largest purity when the size of neighborhood is 1. We found that for SPE the larger the $k$, the better the result. Specifically, when $k$ equals the number of the data, i.e. no filtering at all, SPE turns out to be a nonlinear MDS. As we can see from Figure 3, CTKE and TKE have higher purity than others. The average purity of CTKE is 0.5946 and TKE 0.6135. It shows that CTKE has very close performance to TKE. However, the advantage of the CTKE is its ability to predict novel samples because of the mapping function defined explicitly between input space and latent space. This mechanism broadens the applicability of this algorithm greatly to classification, identification etc. It also provides a tool to explore the manifold formed by data.

6 CONCLUSIONS

In this paper, we visualized the similarity relationships among protein structures using constrained Twin Kernel Embedding (CTKE) which is constructed on the original TKE [10]. It has comparable performance to that of TKE but possesses the ability to predict the embeddings for novel samples. Because CTKE implements a mapping function from input space to latent space, the information among input data is further exploited. CTKE also has the similarity preserving property as TKE does and is purely non-vectorial data applicable since the mapping function comes from the feature mapping and expressed as kernel function eventually. Moreover, there is no $k$-nearest neighbor filtering in CTKE and hence avoid choosing another parameter which is common across other DR algorithms. From the experiments on proteins, we have seen that CTKE preserves the similarity relationships among the protein structures and reproduces them in a much lower dimensional space, allowing easy interpretation by researchers. This algorithm is promising as it can be further applied to the study of the evolution of protein structure and the prediction of proteins functions.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation of China (NSFC 60373090), the ARC DP Development Grant from Charles Sturt University and the University Research Grant from the University of New England.

REFERENCES


