Knowledge Management of 1D SDS PAGE Gel Protein Image Information

Gomathy Ramaswamy[1], Bing Wu[1], Ursula MacEvilly[2]
[1]School of Computing
Dublin Institute of Technology
Kevin Street, Dublin 8, Ireland
ramaswamy.gomathy@gmail.com, bing.wu@dit.ie

[2]School of Biological Sciences
Dublin Institute of Technology
Kevin Street, Dublin 8, Ireland
ursula.macevilly@dit.ie

Journal of Digital Information Management

ABSTRACT: *Computational management and analysis of proteomics data is providing important insights into the biological and clinical sciences. In this research study, a knowledge management strategy for 1D SDS PAGE gel protein image information has been developed and implemented in the New Vision to Protein intensity Analysis System (NeVPAS) application. This application focused on the analysis of skin mucus protein changes during a key developmental stage of Atlantic salmon (Salmo salar), the smoltification period. It includes a database to archive proteomics data. In this paper, knowledge management is expressed in terms of data acquisition, hierarchical organisation of data, knowledge retrieval by means of an information retrieval technique and knowledge presentation using text and graphics. This application of proteomics knowledge management in the marine field may also be extended to clinical analysis and environmental monitoring.*

Categories and Subject Descriptors
H.2.4 [Systems]; Relational databases: **H.3.5 [Online Information Services];** Data sharing, Web-based services**: I.4.1 [Digitisation and Image Capture];** Scanning: **J.3 [Life and Medical Sciences];** Biology and genetics: **K.6.3 [Software Management];** Software development

General Terms: Management, Standardisation, Development

Keywords: Knowledge management, 1D SDS PAGE protein gel image, Gel image analysis, Atlantic salmon

Received: 29th May 2009; Revised: 20th August 2009; Accepted: 23rd September 2009

1. Introduction

A variety of computational supports have been developed to support the rapid generation of biological data from the proteomics domain. The most comprehensive information provided on gel electrophoresis image analysis involves the quantification of protein characteristics on 1D and 2D SDS PAGE gels. Gel electrophoresis along with other analytical techniques such as mass spectrometry and protein-protein interaction network analysis produce large data sets. Knowledge management and discovery in this proteomics domain is a challenging task to computer scientists. Cannataro (2008) described management and analysis of mass spectrometry based proteomics data. The current application domain of this research is ranched Atlantic salmon. The anadromous life cycle of Atlantic salmon undergoes a physiological transformation period called smoltification, in which it prepares for migration from fresh water to sea water. Smoltification status and good physiological condition were shown to be important for the success of Atlantic salmon transfer to the sea (Virtanen et al. 1991). Skin mucus samples of Atlantic salmon were collected from the Marine Institute hatchery in Co. Mayo, Ireland. Protein analysis was carried out by 1D Sodium Dodecyl Sulphate (SDS) Poly Acrylamide Gel Electrophoresis (PAGE) to separate and estimate the molecular weight of proteins within a given skin mucus sample. These gels were digitised and quantified in a standardised manner using an AlphaImager HP, scientific grade digitizing instrument and AlphaEase FC, 1D gel image analysis software. In this research, the research framework (Ramaswamy et al. 2008), the CODES approach (Ramaswamy et al. 2009) have been developed and implemented in an integrated system called NeVPAS, which provides knowledge management of 1D SDS PAGE gel electrophoresis information. The acquired protein knowledge is graphically visualised in the NeVPAS web application. This management of proteomics data can support scientists in the biological and clinical domain to investigate changes occurring in protein expression. In this Atlantic salmon research, management of proteomics knowledge can assist biologists and staff of the marine sector to make decisions regarding the optimum time to release Atlantic salmon to the sea. This paper is divided into four sections: Background; Overview of CODES approach; Knowledge management; and Discussion and Future work.

2. Background

Sternberg (1983) developed a parallel pipelined image processor, called a cytocomputer, and a high-level language specifically created for Biomedical image processing, C-3PL, the cytocomputer parallel picture processing language. He also reported the general consensus of the importance of the application of automated computational analysis of gel electrophoresis images in the medical field. It not only increases diagnostic accuracy but also provides significant data which are not obtainable from ocular analysis. In order to computationally analyse 1D SDS PAGE gel proteins, the protein gels need to be digitised. In the past decade, densitometer and document scanners have commonly been used to scan protein gel images (Lemkin et al. 1999). The applications of other types of scanners such as storage phosphor imagers and CCD cameras were outlined by Ye et al. (1999). Recently software companies have been providing

scientific grade image capturing equipment along with control software to perform scanning more effectively. The authors outlined a standardised camera setup for digitising 1D SDS PAGE gel electrophoresis images (Ramaswamy et al. 2009) using AlphaImager HP. The digitised analysis of electrophoresis gel images emerges as one of the most important applications to reduce human error and increase the speed of data evaluation. To analyse electrophoresis gel images accurately, reliably and quantitatively, protein bands in 1D SDS PAGE gel images which reflect the characteristics of individual bands should be quantified. This can be achieved with computer assistance. Image processing and analysis techniques are capable of extracting the characteristics of protein bands and converting them into numerical data. This can prevent the subjective evaluation of protein information. Ye et al. (1999) developed in-house software for quantifying bands in cDNA gel images.

Software Name	Company	Internet URL
Alpha Ease FC	Alpha Innotech Corporation	http://www.alphainnotech.com/
LabImager 1D 2006 Professional	KAPELAN Bio Imaging Solutions	http://www.labimage.net/
ImageQuant TL	GE Healthcare	http://www.imsupport.com/
GelQuant	DNR BioImaging Systems	http://www.dnr-is.com/
TL120 DM	Nonlinear Dynamics	http://www.nonlinear.com/
1D Main (one dimensional)	BIOTECH software and Hardware	http://www.aabi.com/
GeneTools	Syngene Corporation	http://www.syngene.com/
EZQuant-Gel	EZQUANT Biology software solutions	http://www.ezquant.com/
GelPro	Media Cybernetics	http://www.mediacy.com/
QuantiScan 3.0	BIOSOFT software	http://www.biosoft.com

Table 1. Some of the commercially available software for analysis of 1D SDS PAGE gel images

To date, several commercial software packages have been developed to facilitate rapid, accurate and objective analysis of 1D SDS PAGE gel images. A summary of software available for analysis of 1D SDS PAGE gel images is shown in Table 1. A web based tool called GelScape was reported to provide simple, interactive annotation, manipulation and storage of both 1D SDS PAGE and 2D SDS PAGE protein gel images (Young et al. 2004). GelScape was developed to identify proteins using the amino acid sequence or characteristics of proteins such as molecular weight (MW) and Iso electric point (pI). This information which is stored in a database cannot be shared. GelScape is not intended to perform knowledge management of proteomics expression data. Bromage and Kaattari (2007) used Phoretix 1D gel image analysis software on affinity purified rainbow trout antibody to show the ability of accurate densitometry to conduct quantitative analysis of multiple protein species within a single sample.

In the context of stock enhancement programmes and the conservation of Atlantic salmon stocks, the timely release of salmon from freshwater to sea water is important. Accordingly, this research study sets out to monitor some of the physiological and biochemical changes, which allow Atlantic salmon to migrate from freshwater to seawater. Mucus protein analysis was carried out using 1D SDS-PAGE. Fagan et al. (2003), Kennedy (2004) and Verbeken (2005) observed significant and consistent changes in the 14.5 kDa and 33 kDa regions of the protein profiles. These findings suggested that changes in these mucus protein bands warranted further investigation to determine the physiological significance of the changes observed in relation to the smoltification process. These results were limited to human visual interpretation. Eibrand et al. (2003) developed an application capable of extracting data from the images of 1D SDS PAGE protein gels in JPEG format. This application called COPS incorporated its own image processing and analysis program for 1D SDS PAGE gel protein images. The COPS application identified lanes or tracks in the gel image and protein bands in the lane. The profile plot of a lane was also detailed and an option to quantify the protein bands included. The increasing intensity of a 14.9 kDa protein and decreasing intensity of the protein duplet in the region 33 kDa were confirmed. The COPS application was not suitable for TIFF images. The CODES approach has been developed to analyse and manage 1D SDS PAGE protein information in terms of standardised digitisation, quantification, normalisation, data acquisition, storage, retrieval and visualisation. In this paper knowledge management is concerned with the acquisition, organisation, presentation, use and evolution of proteomics knowledge in many forms. The main goal is to demonstrate the effectiveness and efficiency of the computational system for the management of large quantities of biological data.

3. Overview of CODES approach

The CODES approach, proposed by Dr Bing Wu and Dr Ursula MacEvilly in 2006, was developed (Ramaswamy 2009 ; Ramaswamy et al. 2009) to manage 1D SDS PAGE proteomics information of the skin mucus of ranched Atlantic salmon. Skin mucus samples of Atlantic salmon were taken on three dates (22nd of March, 12th of April and 4th of May) during the smoltification period in 2006. The biologists involved in this interdisciplinary team described a standardised procedure for preparing 1D SDS PAGE protein gels (Verbeken 2005). Three 1D SDS PAGE gel images were prepared from samples of three dates using standard biological procedures. A 1D SDS PAGE gel protein image of skin mucus samples, which were taken on the 4th of May 2006 is shown in Figure 1. The flow diagram of the CODES approach is shown in Figure 2. Image digitisation is the first phase of a computerised approach to analyse images. In recent times, the most common flat bed scanner is not preferred for scanning proteomics gels because of the difficulty in obtaining high quality images and the JPEG image format. The JPEG file format uses a lossy compression method, which discards some image data based on the compression ratio used. Nowadays a number of commercial and professional scientific image capturing devices are available to scan protein gels, which are able to provide

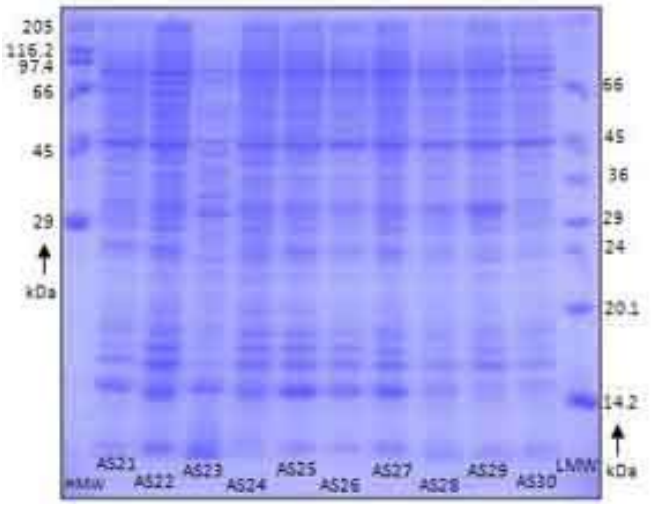


Figure 1. 1D SDS PAGE gel protein image of Atlantic salmon skin mucus samples taken on the 4th of May 2006 (Ramaswamy et al. 2009)

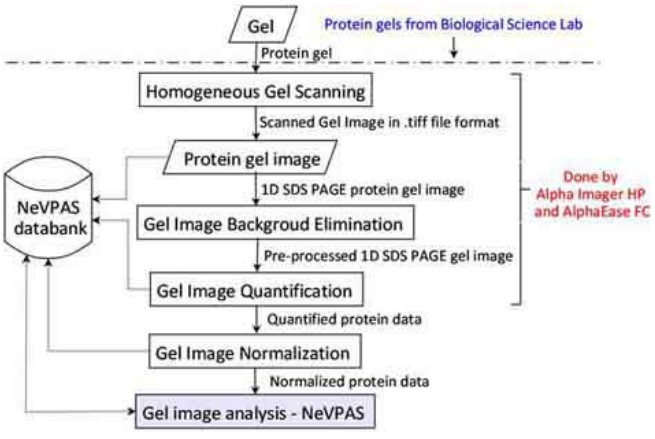


Figure 2. Flow diagram of CODES approach

images in TIFF format. The accuracy of quantification depends on the quality of the digital image.

Hence, it is advantageous to use digital equipment which is capable of giving images which are similar to TIFF images. In this research, the scientific grade scanner AlphaImager HP was used to carry out digitisation. The criteria considered to achieve optimum image quality during the selection of an instrument for image capture are as follows: image type (eg: TIFF, JPEG, GEL), image resolution , light effects (reflecting, trans-illuminating white light / UV light), camera speed and portability. Generally, in proteomics studies comparison are required between multiple gel protein images. Therefore, it is essential to apply standardised image capture settings to every gel protein image involved in the study. Standardisation of image acquisition is the process of establishing a technical image capture standard specification for gel image digitisation.

In this research study, the gel protein image of Atlantic salmon skin mucus sampled on the 22nd of March 2006 was digitised using twenty five different setups. These settings were set to different values of aperture, zoom, focus and exposure time in the camera. Gel images were quantified using AlphaEase FC software for twenty five image capture setups. The protein band intensity variation was determined by examining the intensity of the 36 kDa protein standard. Protein intensities varied between 4,000 and 23,000 pixels for this protein (Figure 3). The 14.2 kDa and 29 kDa standard proteins also examined showed similar

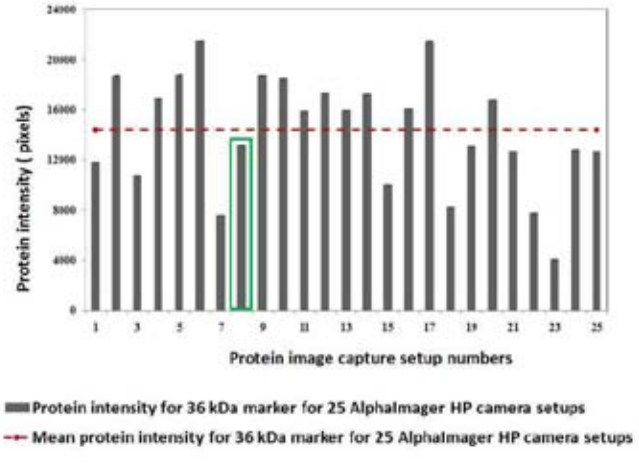


Figure 3. Various image capture sets of the intensities of the 36kDa protein standard on the gel for Atlantic salmon skin mucus sample (LMW, 22nd of March 2006)

variations in intensities. It is evident from these analyses that applying different capture settings to a group of protein gels yields protein gel images which are not appropriate for comparative analysis. It is also evident from Figure 3 that image capture setups 8, 11,13 and 19 are closer to the mean protein intensity value. Images captured by twenty five setups were also visually evaluated and the same four setups (8, 11, 13 and 19) were satisfactory (Table 2). Setup 8 was closest to mean protein intensity and was selected as the standard setup and applied to capture all 1D SDS PAGE gel protein images in the study. Strong intensity changes are visible on protein gels.

Camera setup number	Aperture	Zoom	Focus	Exposure Time
8	**2.0**	**17.50**	**52.00**	**13ms**
19	2.0	15.00	52.00	13ms
13	1.60	17.50	70.80	13ms
11	1.60	17.50	99.00	13ms

Table 2. Suitable camera setups for 1D SDS PAGE gel protein images

Slight intensity changes may not be directly readily detected. Computing technology can provide valuable assistance in detecting and quanlifying such changes. In the quanlification process, extraction of meaningful information from digital images by means of digital image processing techniques was carried out with the support of AlphaEase FC 6.0 software, which was released in the year 2002. The following protein band properties were quantified from the 1D gel image (TIFF format): molecular weight, band intensity, band position and band size, and expressed in numerical terms. This information provided the key resource for the computational phase of this research. The software application called COPS, which was developed by Eibrand et al. (2003) to quantify gel protein bands included some disadvantages in the quantification phase. The protein gels were scanned using a flat bed scanner, which provided protein gel images in JPEG file format. It is known that JPEG images are compressed and not ideally suitable for scientific analysis.

The COPS application was developed using the ImageJ to read only JPEG gel protein images. It detected protein bands and quantified protein intensity. Although the quantification results confirmed the biological findings of Kennedy (2004), the reliability of the software application and the quality of the performance was limited to JPEG images. It initiated further investigations and a scientific grade image capture instrument was obtained which can provide TIFF images. Also commercial gel image analysis software was used for quantifying protein bands in 1D SDS PAGE protein gel images. AlphaEase FC is not specific to 1D SDS PAGE gel protein images. It also contains tools for western blots, micro array analysis and colony counting. In image processing and analysis techniques the intensities of protein bands are important indicators for computational analysis. Generally, the true intensity of a protein band is combined with background intensity. Background elimination techniques were applied to remove unwanted background noise and provide the true intensity of protein bands. The AlphaEase FC software itself contains background removal methods for 1D SDS PAGE gel protein images, namely Auto Base, Minimum Profile, Rubber Band, Intra lane, Rolling Disc and Valley-Valley for background subtraction. The comparison of all of these methods in the AlphaEase FC software for low molecular weight protein markers (14.2 to 66 kDa) are shown

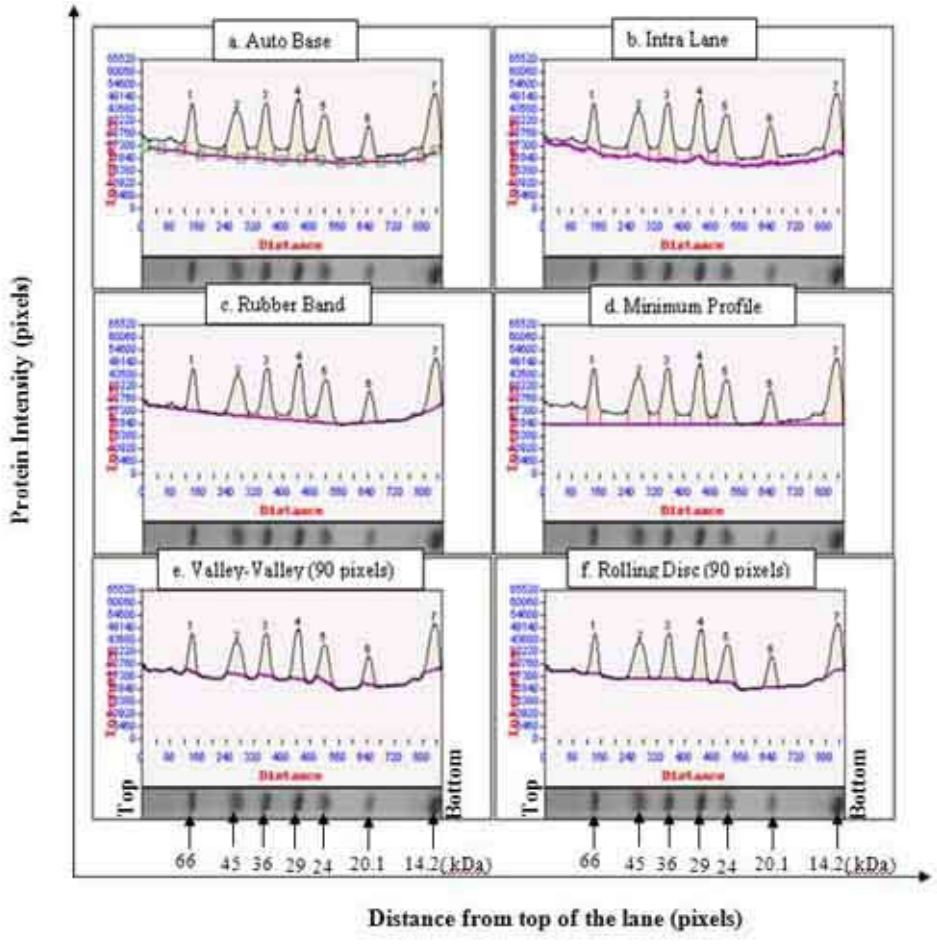


Figure 4. Result of different background subtraction (a – f), Low Molecular Weight (LMW) protein markers on 1D SDS PAGE gel protein image (4th of May 2006)

in Figure 4. The CODES approach used the rolling disc method because it eliminated the background noise present in 1D SDS PAGE gel protein images more effectively.

Gel electrophoresis procedures can be subject to variation due to differences in sample preparation, sample loading, staining and/or destaining time and image acquisition. The aim of normalising gel images is to make them comparable to other gel images. A technical article published by Biostep GmbH described the application of four normalisation methods namely: band percentage, lane percentage, lane normalisation and normalisation to an individual band for 1D gel images and TLC plates (Venning, 2008). These methods were developed to normalise experimental variation between lanes of the same gel image. In this study normalisation is performed to compensate for variation between gel images, instead of lanes in the gel image. It provides comparable gel images and consistency for a set of data, signals, or images. Normalisation of 1D SDS PAGE gel images was performed by comparing, analysing and performing mathematical calculations using the intensities of all the protein markers in the reference gel image and sample gel image. In the normalisation procedure of the CODES approach, one gel image is taken as reference against all other gel images. For example, for images of 1D SDS PAGE protein gels produced from ranched Atlantic salmon skin mucus for the 2006 sample set, the data for the first sampling date (22nd of March 2006) was taken as the reference image in order to compare protein changes occurring in the skin mucus after that date. The data from the other two sampling date images (12th of April 2006 and 4th of May 2006) are sample images. The protein intensity of all molecular weight markers (14.2kDa, 20.1kDa, 24kDa, 29kDa, 36kDa, 45kDa, 66kDa, 97.4kDa, 116kDa and 205kDa) was taken from reference and sample gel images. Differences in intensity between the reference and the sample images were then calculated. The mean difference in intensity was calculated and added to un-normalised protein intensities to yield normalised protein intensity values. The normalisation results for Atlantic salmon skin mucus samples taken on the 12th of April 2006 are shown in Figure 5. The coefficient of variation for intensity of protein markers was reduced by 38 - 89% (Ramaswamy, 2009). This normalised data is now suitable for further analysis.

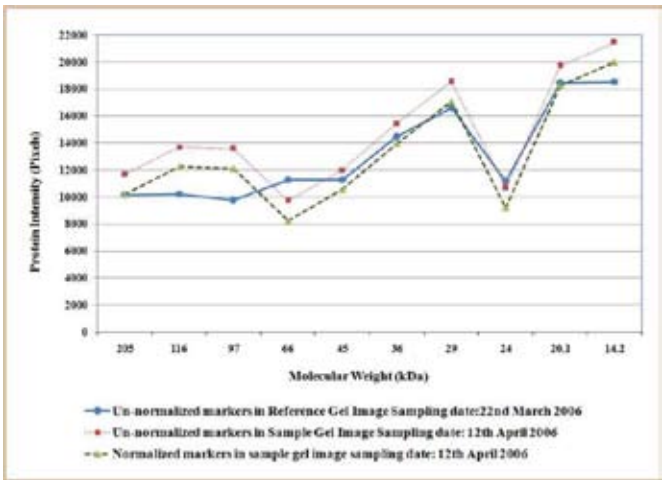


Figure 5. Normalisation plot for Atlantic salmon samples taken on the 12th of April 2006 against the reference gel image on the 22nd of March 2006 (Ramaswamy et al. 2009)

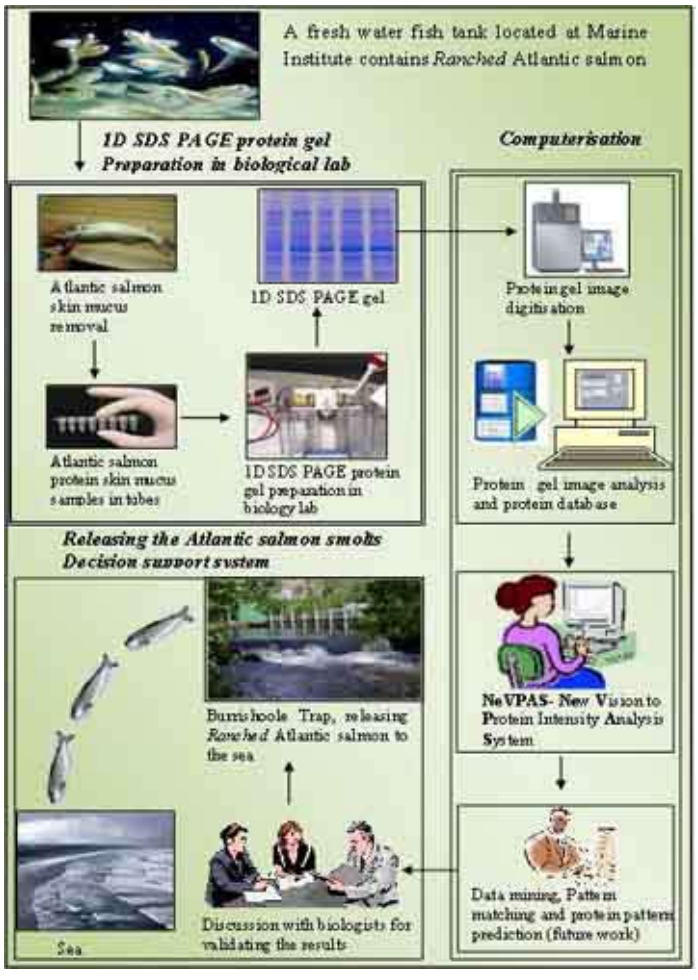


Figure 6. Overview of CODES approach

The importance of the application of computer science to proteomics data is not only quantifying and analysing the protein data but also important in regard to storing, managing and retrieving the digitised gel images, quantified data and analysis results (Maurer et al. 2000). The CODES approach established a databank, named NeVPAS, to store the proteomics studies data. The CODES approach was implemented for analysing and managing protein data by developing the NeVPAS application. It will be useful not only for management and investigation of Atlantic salmon but also suitable for other species. An overview of the CODES approach is shown in Figure 6 (Ramaswamy et al. 2008). This type of approach may be applied to marine, clinical and other biological sectors.

4. Knowledge management

Bellinger (1997) described knowledge is one of the most important resources, which can be effectively used by information analysis experts and computer scientists to assist biologists. Knowledge management is a rapidly expanding theme in which a research team or an enterprise consciously and comprehensively gathers, organises, archives and analyses its knowledge in terms of data, text documents and multimedia documents. It mainly involves data analysis to investigate hidden information and knowledge. The basic unit of knowledge management is data. It is a meaningless point in context and time axis as shown in Figure 7. Information is an understanding of the relationships between pieces of data. Information has a tendency to be relatively static in time and linear in nature. There may be patterns beyond relationships. A pattern which represents knowledge can also provide, when the pattern is understood, a high level of reliability or predictability as to how the pattern will evolve over time. Understanding the principles responsible for the patterns expressing knowledge may come up with wisdom. Wisdom is the most powerful analytical result and can be helpful in prediction, decision making and in improving the quality of the existing system (Bellinger 1997).

4.1 Knowledge management on 1D SDS PAGE proteomics data

The application of computer science to the life sciences requires that some fundamental issues be addressed such as: data acquisition, data preservation, effective information representation, and its

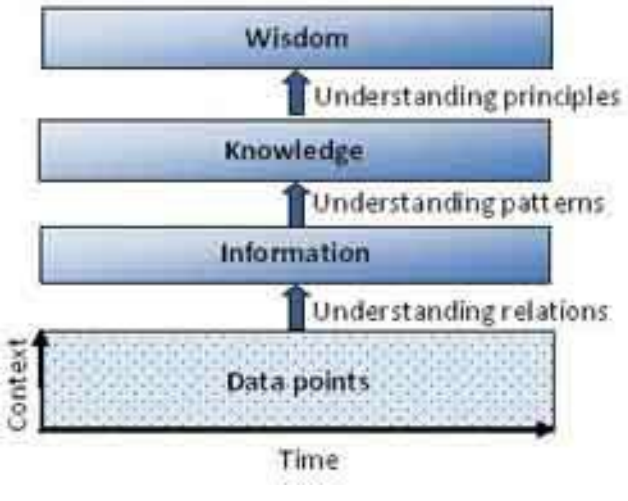


Figure 7. Hierarchy of knowledge management system (Bellinger, 1997)

efficient access, analysis and use during decision making (Jurisica et al. 2005). The systematic knowledge management, analysis and effective use of information can significantly increase the understanding of biology in general and also specific developmental stages and health and disease processes. Computational analysis of data from the proteomics domain requires an iterative and interactive approach, such as CODES, as presentation and organisation of knowledge evolves with an increased understanding of the domain. This requires a new generation of computational applications, such as NeVPAS and their interactive use with the databank and the end user. From the computing perspective, many techniques such as software engineering, databases and information systems were combined for representing and managing protein knowledge. The most common model of data and information representation is a relational data model with the Relational DataBase Management System (RDBMS) (Jurisica et al. 2005).

4.1.1 Data management

RDBMS is a system to store and manage databases with security provisions. A relational database comprises a set of tables, each storing records (or information). A record is represented as a set of attributes (data), which define the property of a record. Designing the conceptual model of RDBMS can be helpful during creation of relational database. It identifies and describes entities, relationships between entities and integrity constraints, which can be represented using entity-relationship diagram (ER) (Jurisica et al. 2005).

In this research, Microsoft SQL Server 2005 database management system has been used to develop the NeVPAS databank. The conceptual model of the NeVPAS databank was designed by considering two requirements of biologists such as data collection and management of proteomics information. The entity relationship diagram of the NeVPAS databank is illustrated in Figure 8. It consists of entities for species, gel, gel image, lane, band and normalised band. The initial step of modelling the species information is to describe primary species information with the help of the domain knowledge expert. This primary information is the key resource for the proteomics domain, therefore it needs to be stored and later can be retrieved to identify the species used in the research study. For example in this CODES research, the Atlantic salmon is the application domain. Information such as weight, height, fish status are collected from the domain expert (Dunne 2008) and stored in the databank with a unique code for each fish. Skin mucus samples of Atlantic salmon were used to prepare 1D SDS PAGE gels. The protein gels are the input to computer science research and should be digitised for further analysis. In the CODES research project AlphaImager HP is used to scan 1D SDS PAGE gel protein images, as discussed in Section 3. The gel entity provides information about important experimental parameters such as protein stain, electrophoresis voltage and current etc, which are specific to that particular gel. Several protein standard samples may be applied to one 1D SDS PAGE protein gel. It is important to specify which samples are involved in a 1D gel and to save this information for long term usage. This type of mapping eliminates redundant entries of species information into gel details. Mapping entities include gel identification number and corresponding fish / species identification numbers. The protein gel image entity in the conceptual model describes the attributes representing the image such as image height, image width, bits per pixel etc.

The 1D SDS PAGE gel image consists of lanes, thus information on these lanes needs to be stored in a separate entity. This is the partial information on a gel image, it comprises number of bands in the lane, a picture of lane image and also a picture of the lane image after detection of bands. The lanes in 1D SDS PAGE gel images contain bands, which are the prime information for proteomics analysis and must be stored in the NeVPAS databank. The band information poses properties such as molecular weight, intensity, etc., which are quantified from AlphaEase FC software. Protein band normalisation produces a new set of data, which will also be stored in the databank. This provides an organised structure in the databank. A survey of public databases revealed that none have reported on Atlantic salmon. The NeVPAS databank has the capability to store 1D

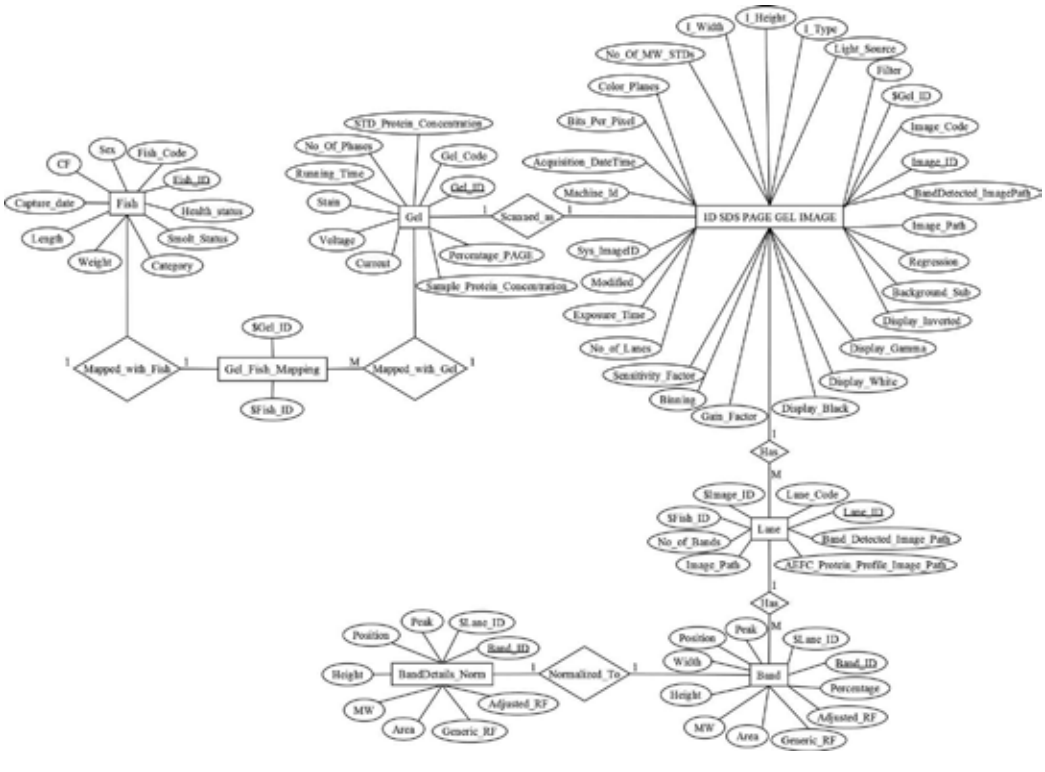


Figure 8. Entity relationship diagram of NeVPAS databank

SDS PAGE protein gel images of all species and other related information in an organised manner. The web forms in the NeVPAS application used Data Definition Language (DDL) and Data Manipulation Language (DML) for creating, modifying and querying data in the databank. The NeVPAS databank applies third normal form to enhance the integrity of data by minimizing redundancy and inconsistency. The data stored is obtained via information retrieval in order to identify the existence of patterns or pattern changes in the developmental stages of species.

4.1.2 Information retrieval

Information retrieval was designed to perform an extensive analysis of 1D SDS PAGE proteomics information to understand the patterns existing in the information. It uses a matching against indices / values technique to fetch and analyse the data. The results of analysis are visualised using tables or graphical aids based on the requirement of biologists. In this research, information retrieval is implemented as the **Ne**w **V**ision to **P**rotein intensity **A**nalysis **S**ystem (NeVPAS) web application. The Microsoft Visual Studio 2005 development suite with Dot Net 2.0 frame work and AJAX were used to develop the NeVPAS application. The information previously stored in the databank can be retrieved as a collection of data. This data was already categorised according to species, gel, gel images, lanes and bands in the databank. Biologists can access and visualise data using reports. These grid reports contain fields which are specific to each category allow access to corresponding tables of relational databank. An image details report (Ramaswamy et al. 2009) of a 1D SDS PAGE gel protein image from the skin mucus of fish sampled on the 4th of May 2006 contains attributes which include image code, gel code, light source, acquisition date and time etc.

It also has hyperlinks to view or download the original TIFF image file and a picture of the image file after band detection. This allows biologists to carry out image processing tasks at any time. Generally, acquisition of proteomics data is a continuing event. This collection of data in a time scale needs to be analysed to find out the existence of patterns, which can yield knowledge. In this research, three sets (six samples in the first group, eight samples in the second group and ten samples in the third group) of data were available from three date time points, the 22nd of March, the 12th of April and the 4th of May 2006. These date time points may or may not be linked. This can be determined by querying the data collection using primary attributes of the protein information such as molecular weight (MW), intensity and date time point. During the process of analysing and understanding the patterns, all samples available at a time point must be considered. It is not only important to analyse and understand associations or patterns from protein information but also necessary to present or visualise the information retrieval results in an effective manner.

The NeVPAS application is intended to perform information retrieval in terms of finding the availability of patterns from a collection of Atlantic salmon protein data and visualizing the results using reports or graphical aids such as bar charts, single line charts and multi line charts. For multiple protein analysis at the same time, this has been implemented in a multi line chart. The NeVPAS application analyses proteomics information by comparing the protein intensity of one or more protein bands on different gel images or different species. In the study to find out the associations between three time points in 2006, protein data analysis can be carried out for any particular protein(s) between all samples. The visual evaluation of changes in proteins for Atlantic salmon skin mucus samples in the 15 kDa and 32 kDa regions were reported by Fagan et al. (2003), Kennedy (2004), Verbeken (2005) and Dunne (2008). Computational analysis set out to quantify the 1D SDS PAGE protein gels using AlphaEase FC software, to store the quantified data in the NeVPAS databank, to find associations between collections of data obtained from skin mucus samples in 2006 and to visualise the associations using the NeVPAS application. The year multi line chart to compare 15, 32, 50, 62 and 85 kDa protein bands for samples collected in 2006 is shown in Figure 9. This is an average pattern of 2006 samples grouped by date time points.

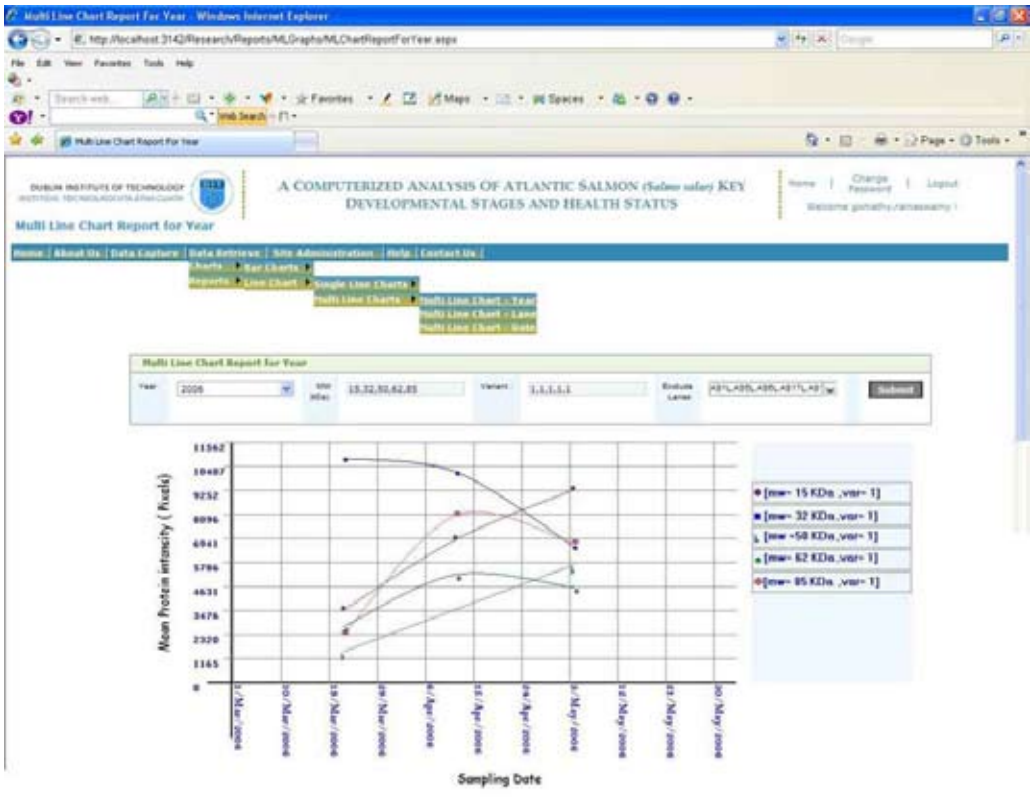


Figure 9. Year multi line chart for the sampling year 2006 for 15, 32, 50, 62 and 85 (± 1 kDa) protein bands

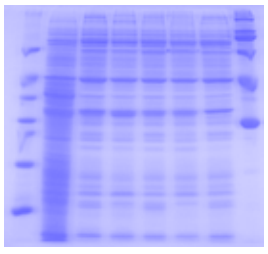
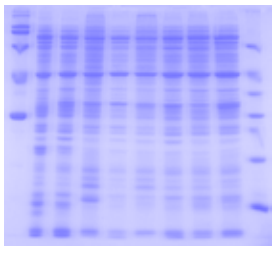
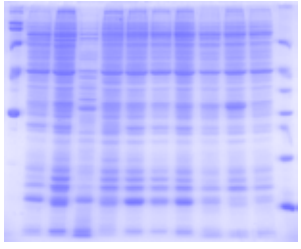
Sampling Date	22nd of March 2006	12th of April 2006	04th of May 2006
Lane Code	AS2- AS4, AS7- AS9	AS11-AS16, AS19, AS20	AS 21 to AS 30
MW Input = 15 kDa			
Output	$AvgInt1 = Int15(\sum_{i=2}^{9} AS_i)$	$AvgInt2 = Int15(\sum_{i=11}^{20} AS_i)$	$AvgInt3 = Int15(\sum_{i=21}^{30} AS_i)$

Figure 10. An example schematic expression for calculating average protein intensity points for molecular weight 15 kDa across samples of year 2006

The schematic expression for calculating average intensities for samples in 2006 is shown in Figure 10. The results confirm an increasing pattern of proteins in the 15 kDa region and decreasing trend in the 32 kDa protein region. The reproducibility and significance of this 15 kDa pattern should be verified with the 15 kDa pattern of other sampling years.

In order to understand the consistency of skin mucus samples taken on one day, the consistency pattern between samples at one time point is also studied. In this research study, ten Atlantic salmon skin mucus samples were collected on the 4th of May 2006. These one day samples may or may not be the same. NeVPAS application has the capability of showing the consistency pattern between multiple samples taken on one day. The consistency pattern for the 15 kDa protein between all samples of the 4th of May 2006 varied in intensity as shown in Figure 11. The physical characteristics of each sample on the 4th of May 2006 are also shown in Figure 12. This data can be further analysed for consistency pattern using data mining techniques. The NeVPAS web application can analyse data from different date points and carry out cross analysis between samples in the databank.

5. Discussion and future work

Computational analysis and management of proteomics information can make a vital contribution to research. In this study, information retrieval for 1D SDS PAGE gel protein image information has been performed with the development of the CODES approach and implementation of the NeVPAS web application. The NeVPAS application is developed as a web based system to evolve a global protein databank for storing 1D SDS PAGE gel images and other related information,

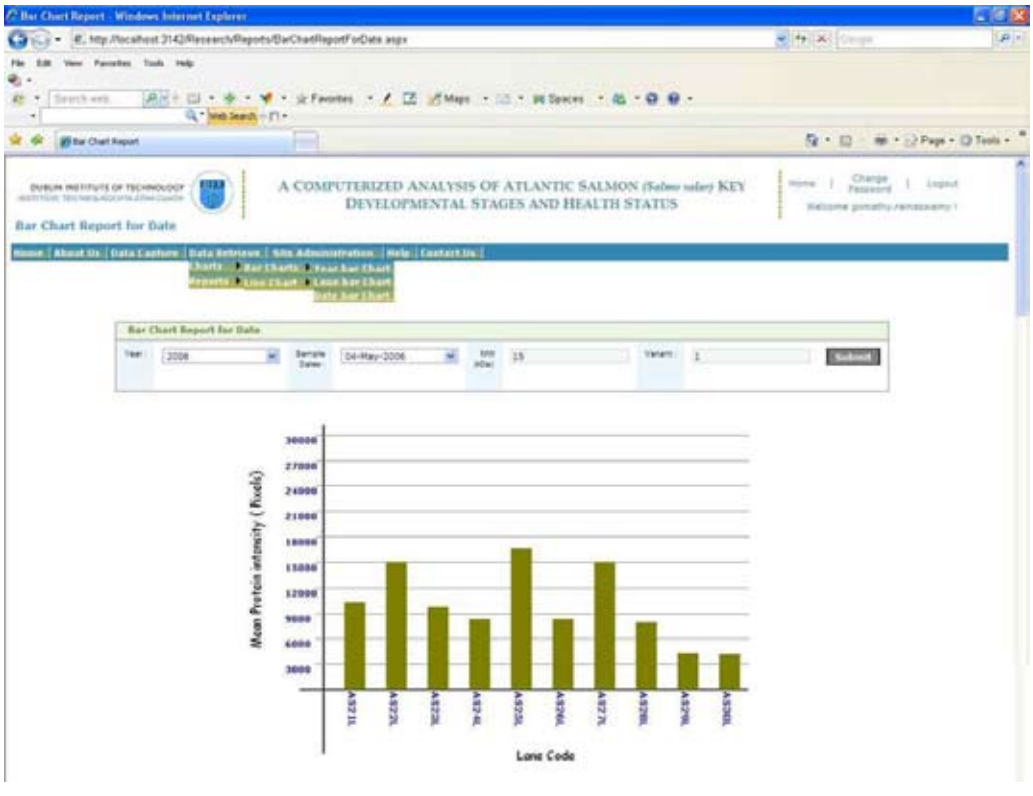


Figure 11. Consistency pattern between samples of the 4th of May 2006 time point at 15 (±1 kDa) protein band

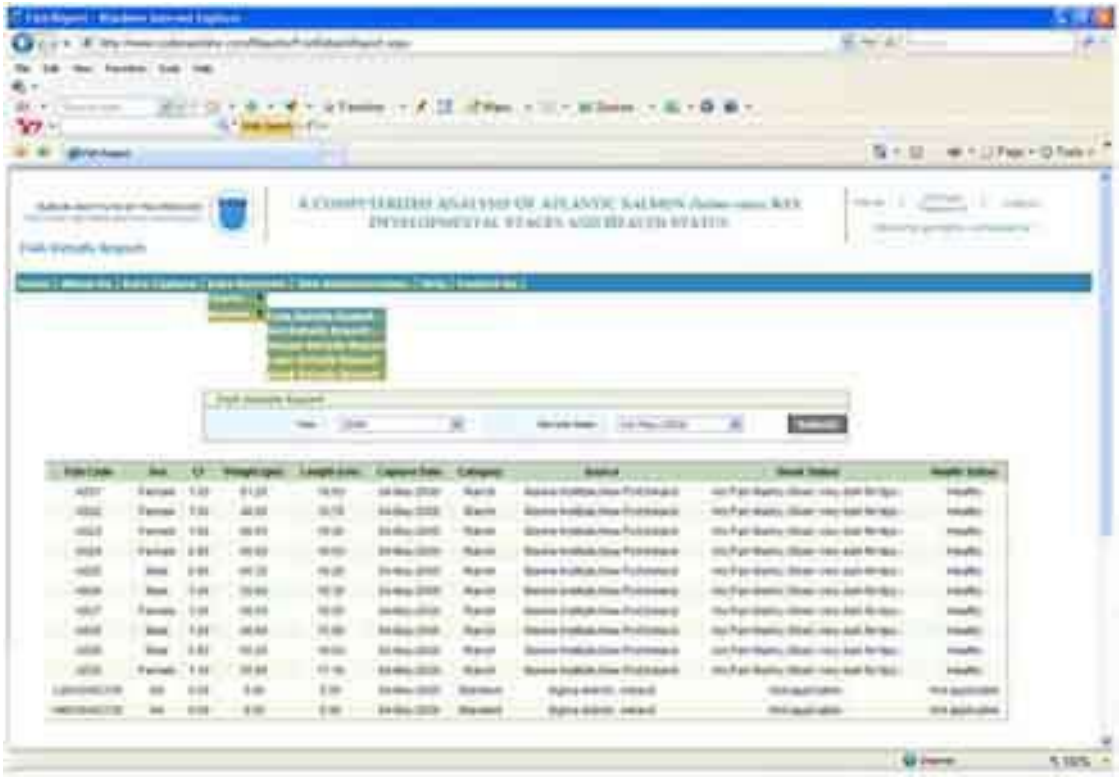


Figure 12. Physical attributes of Atlantic salmon smolts taken on the 4th of May 2006

specifically for Atlantic salmon. The NeVPAS application is a generic suite for information retrieval and facilitates data analysis to investigate pattern(s) in the time domain and to visualise the pattern(s) in an easily accessible way for biologists and medical scientists. It allows researchers to share and compare their results. The NeVPAS application has been evaluated for information retrieval with proteomics data of ranched Atlantic salmon skin mucus samples collected during the smoltification period of 2006 (22nd of March, 12th of April and 4th of May 2006). An understanding of the increasing pattern for 15 kDa protein(s) may provide a greater insight into the developmental stages of Atlantic salmon during the smoltification period. This information retrieval application can be used to determine the consistency, significance, repeatability and reproducibility of patterns in the time domain and can also be used for pattern prediction with additional enhancements. The NeVPAS web application can be enhanced with graphical user interface techniques and statistical reports. NeVPAS databank is the first database developed for Atlantic salmon. It is easily extendable to 2D SDS PAGE gel images. The most popular knowledge management technique ontology can be applied to the NeVPAS application to represent varied concepts, relationships between those concepts and knowledge within a domain (Jurisica et al. 2005). This enhanced, integrated and ontology enabled knowledge management system can accurately assist biologists to study changes which may be important in developmental or disease states of species.

6. Acknowledgements

We thank the Higher Education Authority, Ireland for the funding for this research project. Thanks to Dr Deirdre Cotter (Marine Institute, Newport, Co.Mayo, Ireland) for providing ranched Atlantic salmon samples and Mr Paul Dunne (School of Biological Sciences, Dublin Institute of Technology, Ireland) for preparing 1D SDS PAGE protein gels.

References

[1] Bellinger G. (1997). Knowledge Management. http://www.systems-thinking.org/

[2] Bromage E, Kaattari S. (2007). Simultaneous quantitative analysis of multiple protein species within a single sample using standard scanning densitometry. *Journal of Immunological Methods* 323 (2) 109-113.

[3] Cannataro M. (2008). Computational proteomics: Management and analysis of proteomics data. *Briefings in Bioinformatics* 9 (2) 97-101.

[4] Dunne P. (2008). Proteomics studies of Atlantic salmon (*Salmo salar*) skin mucus proteins. *M.Phil Thesis*, School of Biological Sciences, Dublin Institute of Technology, Dublin 8, Ireland (Expected to publish on 2010).

[5] Eibrand R, Kennedy P, Cotter D, MacEvilly U, Wu B. (2003). Analysis of Atlantic salmon skin mucus: COPS-a computer-based system for protein pattern analysis of 1D SDS-PAGE gels. *In: Proceedings of the* 3rd *IEEE International Symposium on BioInformatics and BioEngineering (BIBE 2003)*, p. 45-52. IEEE, Mar.

[6] Fagan MS, O'Byrne-Ring N, Ryan R, Cotter D, Whelan K, MacEvilly U. (2003). A biochemical study of mucus lysozyme, proteins and plasma thyroxine of Atlantic salmon (*Salmo salar*) during smoltification. *Aquaculture* 222 (1) 287-300.

[7] Jurisica I, Wigle D. (2005). Knowledge discovery in proteomics. CRC Press, ISBN-13: 978-1584884392, p. 360.

[8] Kennedy P. (2004). A biochemical study of Atlantic salmon (*Salmo salar*) skin mucus proteins during smoltification (1999 and 2001). *M.Phil Thesis*, School of Biological Sciences, Dublin Institute of Technology, Dublin, Ireland.

[9] Lemkin PF, Myrick JM, Lakshmanan Y, Shue MJ, Patrick JL, Hornbeck PV, Thornwal GC, Partin AW. (1999). Exploratory

data analysis groupware for qualitative and quantitative electrophoretic gel analysis over the Internet-WebGel. *Electrophoresis* 20 (18) 3492-3507.

[10] Maurer SM, Firestone RB, Scriver CR. (2000). Science's neglected legacy. *Nature* 405 (6783) 117-120.

[11] Ramaswamy G. (2009). A computerised approach to the management of 1D SDS PAGE gel image protein information during the smoltification stage of Atlantic salmon (*Salmo salar*). *M.Phil Thesis*, School of Computing, Dublin Institute of Technology, Dublin 8, Ireland.

[12] Ramaswamy G, Dunne P, Cotter D, MacEvilly U, Brankin B, Wu B. (2008). A computerized analysis of Atlantic salmon (*Salmo salar*) developmental status. *In: Proceedings of the* 3rd *International Symposium on Health Informatics and Bio Informatics (HIBIT 2008),* May.

[13] Ramaswamy G, Wu B, MacEvilly U. (2009). CODES: A computerised approach to the analysis and management of 1D SDS PAGE gel protein image information. *In: Proceedings of the 2nd International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2009)*, p. 578-583. IEEE, Aug. 2009.

[14] Sternberg S. (1983). Biomedical image-processing. *Computer* 16 (1) 22-34.

[15] Venning B. (2008). Electrophoresis Band Quantification-An Art or A Science, Biostep GmbH, Meinersdorfer Straße 47a, 09387 Jahnsdorf, Germany.

[16] Verbeken K. (2005). Biochemical studies of Atlantic salmon (*Salmo salar*) skin mucus proteins during smoltification(2003 and 2004). *M.Phil Thesis*, School of Biological Sciences, Dublin Institute of Technology, Dublin, Ireland.

[17] Virtanen E, Söderholm-Tana L, Soivio A, Foreman L, Muona M. (1991). Effect of physiological condition and smoltification status at smolt release on subsequent catches of adult salmon. *Aquaculture* 97 (2-3) 231-257.

[18] Ye X, Suen CY, Cheriet M, Wang E. (1999). A recent development in image analysis of electrophoresis gels. *In: Proceedings of the Conference on Vision Interface (Vision Interface'99)*, p. 432-438. May, 1999.

[19] Young N, Chang Z, Wishart DS. (2004). GelScape: a web-based server for interactively annotating, manipulating, comparing and archiving 1D and 2D gel images. *Bioinformatics* 20 (6) 976-978.

Authors Biographies

Mrs Gomathy Ramaswamy completed a two year M.Phil research program at School of Computing, Dublin Institute of Technology (DIT), Ireland in June 2009. Worked as Lecturer at Noorul Islam College of Engineering, India for three years during 2003 - 2006. Graduated from the Department of Computer Science and Engineering, National Engineering College, Tamil Nadu, India with an M.E (Computer Science and Engineering) degree in 2003. Research interests include Image Processing and Analysis, Data Mining, Pattern Matching, Pattern Prediction and Artificial Neural Network.

Dr Bing Wu is currently the Head of the Computer Science Department at the Dublin Institute of Technology (DIT), Ireland. Received BSc and MSc in China, and PhD from the University of Manchester Institute of Science and Technology (UMIST), UK in 1996. Worked in the Department of Computer Science, Trinity College Dublin during 1996-1998 and joined DIT in 1998. Research interests include knowledge representation and management, system engineering and re-engineering, and application of advanced computing techniques to real world, such as healthcare informatics and bio-informatics, as well as industry-oriented software education.

Dr Ursula MacEvilly Assistant Head, School of Biological Sciences, DIT 2003 until retirement August 2008. Lecturer in Biotechnology and Microbiology 1984-2003. DIT Co-ordinator EU projects 2000 - 2009. Research interests during recent years included biochemical investigations of key developmental stages and environmental and parasitic stresses on skin mucus proteins of Atlantic salmon, and the development of computerized proteomics systems for protein pattern analysis of fish skin mucus proteins. Previously held quality management and research and development positions in the pharmaceutical and beverage sectors.