Elektrotehniški vestnik 74(3): 119–124, 2007
Electrotechnical Review, Ljubljana, Slovenija

Common open representation of computer vision results in 2DGE research

Peter Peer[1,2], **Victor Segura**[1], **Franc Solina**[2]

[1] *CEIT and Tecnun (University of Navarra), Manuel de Lardizabal 15, 20018 San Sebastian, Spain*
[2] *University of Ljubljana, Faculty of CIS, Tržaška 25, 1000, Ljubljana, Slovenia*
E-mail: peter.peer@fri.uni-lj.si

Abstract. Data standards are required to enable the development of public data repositories, to improve data sharing and also to enable sound and objective evaluation of different approaches. Here, we introduce a flexible data-format for computer vision results in 2DGE (2D Gel Electrophoresis), which meets data standard expectations. This data-format presents an attempt to standardize the communication between research groups in proteomics and computer vision communities, enabling them to share the data and simply compare automatically obtained image processing results of different computer vision algorithms. The proposed CVRML (Computer Vision Results Markup Language) for 2DGE format is modeled in UML (Unified Modeling Language) and implemented in XML (eXtensible Markup Language). Software is being developed, which helps us to easily and effectively annotate and visualize the data. We also discuss a possible framework around the annotated database structured in the CVRML manner, which we intend to make publicly available. All the material described in the paper is freely available on the CVRML web portal.

Key words: image processing, algorithm, 2DGE, comparison, data sharing

Predlog standarda za predstavitev rezultatov procesiranja 2DGE slik

Povzetek. V članku predstavljamo fleksibilen podatkovni format za predstavitev rezultatov procesiranja 2DGE (2D Gel Electrophoresis) slik. Format predstavlja predlog standardizacije komunikacije med raziskovalnimi skupinami v proteomiki in računalniškem vidu. Raziskovalcem omogoča deljenje podatkov in preprosto primerjavo rezultatov različnih algoritmov računalniškega vida. Predlagani jezik CVRML (Computer Vision Results Markup Language) je modeliran v UML (Unified Modeling Language) in implementiran v XML (eXtensible Markup Language).

Ključne besede: procesiranje slik, algoritem, 2DGE, primerjava, deljenje podatkov

1 Introduction

Computer vision is a research line which tries to extract as much information from images as possible. Medical image analysis continues to be an active area of research, with many encouraging results, but also with a number of difficult problems still to be addressed [1].

Proteomics [2] is defined as the systematic large-scale analysis of protein expression under normal and perturbed (stressed, diseased, and/or drugged) states, and generally involves the separation, identification, and characterization of all of the proteins in a cell or tissue sample. There is a broad range of technologies used in proteomics, but the central paradigm has been the use of 2D gel electrophoresis (2DGE) [2, 3] followed by mass spectrometry. 2DGE is used to first separate the proteins by isoelectric point and then by weight.

A variety of commercial medical imaging equipment now comes loaded with simple forms of image processing and analysis algorithms. If we focus on electrophoresis, such systems for example are PDQuest (http://www.bio-rad.com/), Melanie (http://www.expasy.org/melanie/) and DeCyder (http://www.amershambiosciences.com/). These tools nowadays also help us to identify spots of interest for mass spectrometry, but to get there a user has to (very) actively participate in the process. This is especially true for standard 2DGE images, which are of our primary interest; but, as you will see, our data-format could also be used for 2D

Received 29 August 2006
Accepted 23 March 2007

DiGE (Difference Gel Electrophoresis) experiments. On the other hand, computer vision community is mainly interested in completely automatic processing of 2DGE images and objective comparison of results of different approaches.

Computer vision systems are penetrating into a number of areas, since we all wish to have systems that automatically and effectively grasp needed information from images. Proportionately to that there is a lot of different methods and results. But there is a lack of a suitable denominator for presentation of results. Namely, there are many systems which solve similar problems (e.g. alignment of two images), even on different levels (e.g. spot detection only). Unfortunately, integration or comparison of results of different systems is difficult, also because of un-unified data-format for presenting results. We wish to be able to integrate results of different systems and levels of computer vision in a standard but flexible way. In this way we could, besides the images themselves, give the actual contents of the image (the ground truth) and gradually add analyses results of different systems. This would bring easier comparison of results and easier integration of different levels of image processing.

Thus, our main motivation is the formation of a common database, where algorithms can be compared and contrasted to each other using a common XML-based [4] (http://www.w3.org/TR/2004/REC-xml-20040204/) data-format for computer vision results. Both communities will benefit from this data-format: the computer vision community will be able to objectively contrast the efficiency of different approaches, the proteomics community will be able to objectively select the method, i.e. the software package that best suits their needs and the data will easily flow between the communities. Thus, the annotated database of 2DGE images and experiments will be publicly available. To our knowledge, this idea is novel, even in a wider context of only computer vision, as we have not found references that describe similar goals.

One of the major challenges for the computer vision field is the one that remains a critical issue in terms of all practical and theoretical development: the need to develop appropriate validation and evaluation approaches. This challenge has a variety of aspects to it, some of which people in the field have been trying to address. The first is the formation of common databases, where algorithms can be compared and contrasted to each other. The literature remains full of papers that evaluate algorithms on a few trial datasets from the home institution. This is in part due to simple lack of availability of a test set. A very small set of test databases are beginning to be made available to the community. However, much more needs to be done in setting up such test databases, and funding agencies and review panels need to take this more seriously. The second issue in evaluation is the need to not only develop databases, but to cultivate research that will focus on the development of evaluation methodology. We as a community cannot continue to cry out for better validation/evaluation and then not embrace such efforts more fully in our review panels and literature.

At the moment we are establishing a publicly available database using a flexible data-format for computer vision results in proteomics – Computer Vision Results Markup Language (CVRML) for 2DGE – presented in this paper. This format enables a simple comparison of results of algorithms to ground truth information, comparison between different algorithms (also developed by different groups), and a simple way to share the data.

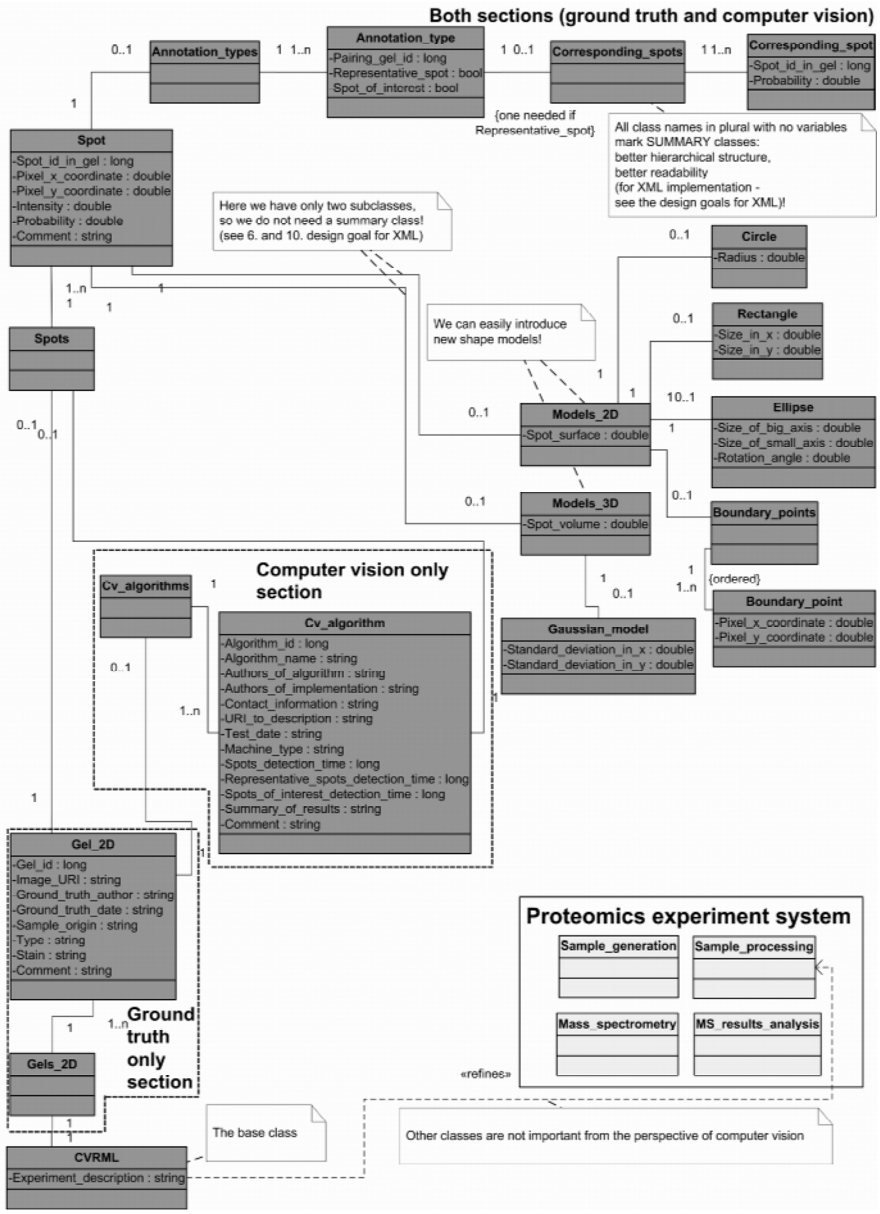
When we were shaping our conceptual model, we were focusing on development of a practical model for 2DGE and not on as general a model as possible. The fact is that a very complex analysis can introduce general but also unpractical solution. The model is intended to act primarily as a spur to discussion about what a final version of such a model might look like, but it represents a functional solution to the design problem.

2 CVRML conceptual model

We developed our conceptual model in UML [5] (Unified Modeling Language: http://www.uml.org/), which is in general a set of diagramming techniques designed to improve software engineering and requirements capture. UML is an industry standard object-oriented modeling language. In this context, it also allows us to describe experimental methods, results, and subsequent analyses in an implementation-independent manner.

Figure 1 shows the CVRML UML class diagram which provides a conceptual model of how to integrate automatically obtained computer vision algorithm results with existing data about one 2DGE experiment. It presents the basis for the XML implementation, revealing all the relations between classes and variables. Namely, each box in a diagram represents a class, which in the XML implementation becomes a tag, while their variables become subtags.

On the lower right side of the diagram we can see basic components of the whole proteomics experiment system. Everything starts with sample generation, continues with sample processing, mass spectrometry and ends with analysis of mass spectrometry results. Our proposed model refines the second



All class names in plural with no variables mark SUMMARY classes: better hierarchical structure, better readability (for XML implementation - see the design goals for XML)!

Here we have only two subclasses, so we do not need a summary class! (see 6. and 10. design goal for XML)

We can easily introduce new shape models!

The base class

Other classes are not important from the perspective of computer vision

Figure 1. CVRML conceptual model presented as the UML class diagram. See text for in-depth explanation.

step in this system, i.e. the sample processing step.

The base class is called ‘CVRML’ and it encompasses all important information about the experiment from the perspective of computer vision. As such it can be used to identify which information is needed if someone would like to extend existing proteomics or 2DGE experiment models with the CVRML idea. Two extensions, extension of the PEDRo (Proteomics Experiment Data Repository) [6, 7] model and the AGML (Annotated Gel Markup Language) [8, 9] model, are presented on the CVRML web portal (see URL at end of Discussion).

The only variable inside this class is *Experiment_description*, which reveals basic logical relations between the gels in the experiment, the hypothesis behind it, etc. Physical relations are visible on a lower level, i.e. when we relate the spots between each other.

Our main goal when we were building the model was to present a flexible data-format for computer vision results in 2DGE. Therefore, having in mind the flexibility demand, XML was a natural choice as it is presently the consensus choice in most areas. We explain details about the implementation of the model in the XML in the next section, but here we have to explain a term, which we named summary class. In our case all classes that have names in plural and are without variables mark summary classes. From the term itself it follows that such a class summarizes a number of instances of the same subclass. This concept is supported by the XML design goals (http://www.w3.org/TR/2004/REC-xml-20040204/#sec-origin-goals), which also emphasize the importance of human-legibility of XML files. And our hierarchical structure with the concept of summary classes significantly improves the readability. This is especially notable if the generated files are big, which is quite common in proteomics. This concept can also be found in the AGML. In our model a summary class is always associated with at least one instance of the subclass.

The first summary class in the model is 'Gels_2D', which contains all the gels associated with the experiment. Inside it there has to be at least one gel (note that the relation between the two classes is 1 to 1..n). Each gel is represented as an instance of class 'Gel_2D'. In contrast to the AGML file, the CVRML file and consequently the 2DGE experiment described in the file can consist of gels of different samples and even of different sample origins, as also suggested in the PEDRo. Since 'Gel_2D' class is common to all known models (e.g. 'Gel2D' class in PEDRo, 'real' class in AGML and '2D-PAGE' in Gla-PSI (Glasgow – Proteomics Standards Initiative) [10, 11]) it is quite straightforward to integrate the CVRML idea into each of them.

The detailed description of the model, classes and variables can be found on the CVRML web portal (see URL at end of Discussion).

3 CVRML implementation

As mentioned above, our goal is to present flexible XML-based data-format for computer vision results in 2DGE. That is why the next step was the implementation of the model in the XML Schema Definition (XSD) language (http://www.w3.org/TR/xmlschema-0/). The XSD enables you to define the structure and data types for XML documents. An XML Schema consists of components such as type definitions and element declarations. These can be used to assess the validity of well-formed element and attribute information items, and furthermore may specify augmentations to those items and their descendants. The XML document is valid if an element or attribute information item satisfies the constraints embodied in the relevant components of an XML Schema. In other words, the schema defines the grammer and the XML documents that follow its rules are nothing more than words in language defined by the schema.

As such the schema gives more details about the proposed data-format in comparison to the conceptual model. And it can be used to verify the validity of generated XML documents.

An implementation decision was taken to follow a recommendation that all tags be elements with no attributes (http://www.oasis-open.org/cover/elementsAndAttrs.html). Furthermore, to ensure good readability of our data-format, even if it gets integrated into more general data-format about the proteomics experiment, we defined our own XML namespace (cvrml) (http://www.w3.org/TR/1999/REC-xml-names-19990114/). It helps us to identify our tags, since in general the tag names can repeat themselves.

Our schema is available on our web portal (see URL at end of Discussion): At the moment the variable *Experiment_description* in the base class is optional. Next, only the first two variables in class 'Gel_2D' are mandatory, because inside the whole XML file we have to know the identifier of each gel and its location, since the gel image is the basic source of information for each computer vision algorithm. In class 'Cv_algorithm' the mandatory variables are *Algorithm_id*, *Authors_of_algorithm*, *Authors_of_implementation*, *Contact_information* and *URI_to_description*. Each spot must have its identifier, the location of the center and its intensity (the first four variables in class 'Spot'). Figure 2 clearly illustrates the implementation of the 'Annotation_type' class. It follows from the definitions that a spot cannot be marked as a representative spot and as a spot of interest at the same time. Thus, we are talking about exclusive or operation. In the figure this fact is clearly presented: we have to make a choice between two options. All the dashed boxes represent optional information. Finally, in all possible spot models ('Circle', 'Rectangle', 'Ellipse', 'Boundary_point' and 'Gaussian_model') all variables are mandatory.

4 CVRML usage

By using the CVRML schema, we can verify the validity of the generated XML file, i.e. we get an answer whether it follows the given grammar, the CVRML idea (e.g. by using XMLSpy: http://www.altova.com/products_ide.html). On the CVRML web portal (see URL at end of Discussion) there is also an exemplary XML file that conforms to the schema. Beside it there are two stylesheets, which display the example as HTML, with a different level of details. Note that, by applying a predefined transformation expressed in XSLT (eXtensible Stylesheet Language Transformation is a language for transforming XML documents into other XML docu-

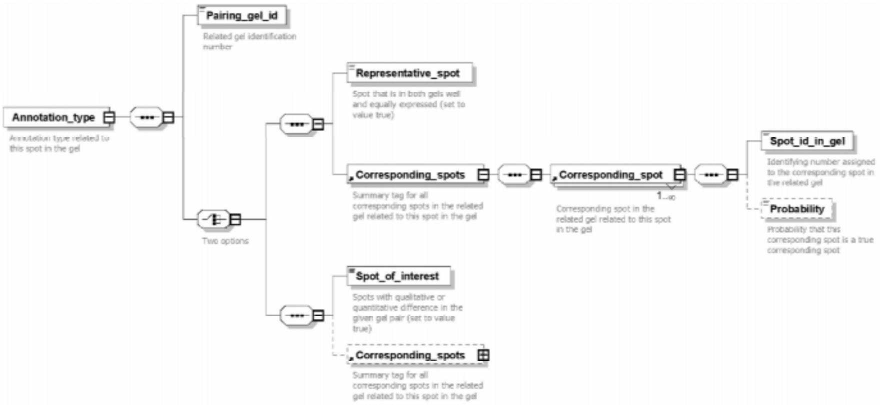

Pairing_gel_id
Related gel identification number
Annotation_type
Annotation type related to this spot in the gel
Representative_spot
Spot that is in both gels well and equally expressed (set to value true)
Corresponding_spots
Summary tag for all corresponding spots in the related gel related to this spot in the gel
Corresponding_spot
1..∞
Corresponding spot in the related gel related to this spot in the gel
Spot_id_in_gel
Identifying number assigned to the corresponding spot in the related gel
Probability
Probability that this corresponding spot is a true corresponding spot
Two options
Spot_of_interest
Spots with qualitative or quantitative difference in the given gel pair (set to value true)
Corresponding_spots
Summary tag for all corresponding spots in the related gel related to this spot in the gel

Figure 2. Illustration of the implementation of the 'Annotation_type' class. See text for details.

ments: http://www.w3.org/TR/xslt), the XML files can be read directly, as HTML, using a normal web browser. This is very useful when we want to get a general overview about the 2DGE experiment and computer vision results and conclusions.

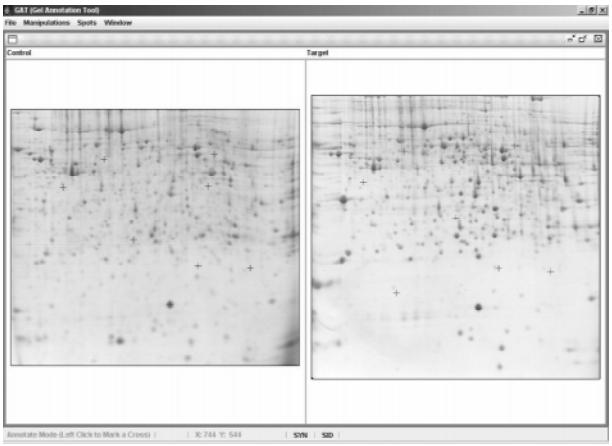

GAT (Gel Annotation Tool)
File
Manipulations
Spots
Window
Control
Target
Annotate Mode (Left Click to Mark a Cross)
X: 744 Y: 544
SYN
SID

Figure 3. Basic graphical user interface of GAT (Gel Annotation Tool) application, our tool which helps us to easily and effectively annotate the data in the CVRML manner.

A repository that would conform to the CVRML idea requires a tool to manage the data. Figure 3 shows a basic graphical user interface of GAT (Gel Annotation Tool) application, our tool under development, which helps us to easily and effectively annotate the data in the CVRML manner. In GAT we identified a few basic use cases: load gels related to one experiment, manipulate images (e.g. zoom, print etc.), annotate gels, save and load annotations and control the level of details displayed about each gel. At the moment we are also working on image analysis module for proteomic experiments which use 2DGE. We are developing an algorithm that will automatically annotate data and reduce the amount of user interaction, if not completely eliminate it, during the whole image analysis process. To effectively and objectively test our algorithm, we have to have a repository of ground truth and computer vision annotated gel images. Our database of images is getting bigger and now that we have a proper data-format, we will be able to give the database together with annotation application to the proteomics and computer vision communities. Our primary target of course are computer vision researchers researching in the field of 2DGE, but our data-format also enables better communication between the communities. In our case it represents a bridge between people with different backgrounds: biology, informatics and mathematics.

Figure 4 shows one possible framework around the database. It incorporates all the presented ideas into one system which describes the flow of information. All parts of the framework except the box named Analysis have already been described. This box represents a tool which enables the researchers to contrast the ground truth information to computer vision results and also to compare computer vision algorithms among each other. This functionality could of course also be a new use case in GAT.

5 Discussion

Our intention in this paper was to describe a flexible data-format for computer vision results in 2DGE. We started with a conceptual model of our CVRML markup language for 2DGE, then we introduced its implementation and usage, pointing out also our plans for future work. Our format enables a simple comparison of results of computer vision algorithms to ground truth information, comparison between different algorithms and a simple way to share the data. As such it represents a functional solution to the design problem.

The CVRML provides a common language for proteomics and computer vision communities, which

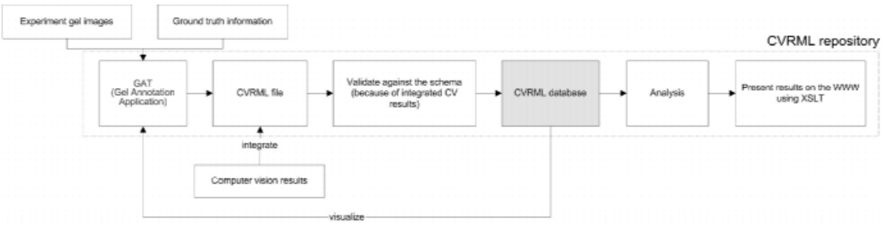


Figure 4. Possible framework around the CVRML database.

means that with a proper set of free tools to work with it, the proposed format will also facilitate the exchange of data, promoting collaboration and creation of public data repository [12]. Since CVRML conforms to the XML rules, it's highly flexible and simple to modify [13].

The data presented in the CVRML file is never redundant, keeping its size to the minimum. Nevertheless, all the relations between its parts are integrated. That also means that we can easily for example present all spots of interest in one experiment (that consists of many gels) in a single, virtual gel image.

The URL at the end of this section will take you to the CVRML web portal, where all the material presented in this paper and more is available. On it you will find our conceptual model, detailed description of the model and its integration into PEDRo and AGML. Then there is the implementation of the model in the XML Schema Definition (XSD) language, an exemplary XML file that conforms to the schema and two stylesheets, which display the example as HTML, with a different level of details. The next set of files available on the portal is devoted to the documentation of the schema. There is a graphical representation of hierarchical relations between the tags, general description of tags and detailed description of tags with the associated and explained source code.

URL

The CVRML web portal: http://www.lrv.fri.uni-lj.si/~peterp/CVRML/CVRML.htm.

Acknowledgments

Special thanks go to Elizabeth Guruceaga, Angel Rubio, and Fernando J. Corrales for valuable discussions about the CVRML data-format and its application. This work was supported by the Department of Education, Universities and Research of Basque Government and the Ministry of Education, Science and Sport of the Republic of Slovenia.

6 References

[1] Duncan, J.S. & Ayache, N. Medical image analysis: progress over two decades and the challenges ahead. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 85–106 (2000).

[2] Patterson, S.D. & Aebersold, R.H. Proteomics: the first decade and beyond. *Nat. Genet.* **33**, 311–323 (2003).

[3] Ong, S.-E. & Pandey, A. An evaluation of the use of two-dimensional gel electrophoresis in proteomics. *Biomol. Eng.* **18**, 195–205 (2001).

[4] Simpson, J.E., *Just XML* (Prentice Hall PTR, NJ, USA, 2001).

[5] Rumbaugh, J., Jacobson, I. & Booch, G. *The Unified Modeling Language reference manual* (Addison-Wesley, Reading, MA, USA, 1999).

[6] Taylor, C.F. et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* **21**, 247–254 (2003).

[7] Garwood, K. et al. PEDRo: A database for storing, searching and disseminating experimental proteomics data. *BMC Genomics* **5** (2004).

[8] Stanislaus, R., Jiang, L. H., Swartz, M., Arthur, J. & Almeida J. S. An XML standard for the dissemination of annotated 2D gel electrophoresis data complemented with mass spectrometry results. *BMC Bioinformatics* **5** **(**2004).

[9] Stanislaus, R., Chen, C., Franklin, J., Arthur, J. & Almeida J.S. AGML central: web based gel proteomic infrastructure. *Bioinformatics* **21,** 1754–1757 (2005).

[10] Jones, A., Wastling, J. & Hunt, E. Proposal for a standard representation of two-dimensional gel electrophoresis data. *Comp. Funct. Genom.* **4**, 492–501 (2003).

[11] Jones, A., Hunt, E., Wastling, J.M., Pizarro, A. & Stoeckert, C.J. Jr. An object model and database for functional genomics. *Bioinformatics* **20**, 1583–1590 (2004).

[12] Prince, J.T., Carlson, M.W., Wang, R., Lu, P. & Marcotte, E.M. The need for a public proteomics repository. *Nat. Biotechnol.* **22**, 471–472 (2004).

[13] Achard, F., Vaysseix, G. & Barillot E. XML, Bioinformatics and data integration. *Bioinformatics* **17**, 115–125 (2001).

Peter Peer, PhD, is an assistant professor at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. His research is focused on computer vision.

Victor Segura is a PhD student at CEIT and Tecnun (University of Navarra), Spain. His research is focused on proteomics and genomics in relation with computer science.

Franc Solina, PhD, is a full professor at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. His research is focused on computer vision.