Periodicals of
Engineering and Natural Sciences

Periodicals of Engineering and Natural Sciences (PEN)

Periodicals of Engineering and Natural Sciences (ISSN: 2303-4521) is an international open access single-blind review journal published online.

Publication frequency: Semiyearly (1. January - June; 2. July - December).

Publication Fees: No fee required (no article submission charges or processing charges).

Digital Object Identifier DOI: 10.21533/pen

Editorial Team

Publication Frequency

Expected frequency of publication is twice per year:

1. January - June;

2. July - December.

Journal items can be published collectively, as part of an issue with its own Table of Contents. Alternatively, individual items can be published as soon as they are ready, by adding them to the "current" volume's Table of Contents.

Open Access Policy

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

Journal Ethics and Malpractice Statement

PEN Journal is committed to ensure and uphold standards of ethical behavior at all stages of the publication process, whereby such values also rely on editors, reviewers, contributors and authors who are expected to behave ethically. The standards are based on Committee on Publication Ethics' (COPE) code of conduct, and provide guidelines for best practices in order to meet these requirements. The following ethical guidelines are only intended to give a summary of our key expectations of editors, peer-reviewers, and authors but if you have any questions or concerns please also feel free to contact the Editor of the Journal.

1. Ethical Expectations

Editors' responsibilities

- To carry out their duties in a fair, objective and consistent manner, without discrimination on grounds of gender, sexual orientation, religious or political beliefs, ethnic or geographical origin of the authors.
- To take care that each article undergoes the proper peer-review process.
- To promote consistent ethical policies of PEN journal, and to ensure the confidentiality of the reviewing process.
- To uphold integrity in their work as editors of the journal, doing away with any personal interest, so that articles are considered and accepted solely on their academic merit and without commercial influence.
- To work with authors, reviewers, and the Editorial Board members to ensure the implementation of journals' ethics and publishing policies.
- To adopt and follow reasonable procedures in the event of complaints of an ethical or conflict nature, in accordance with the policies and procedures. To handle process of complaints and giving authors an opportunity to respond any complaints. All complaints should be investigated and the documentation associated with any such complaints should be retained.

Reviewers' responsibilities

- To act objectively, fairly and in a timely manner in reviewing submitted manuscript with the aim of improving its quality such as pointing out relevant published work, which is not cited, etc.
- To alert PEN in case of any competing interest that could affect the impartiality of their reviewing, or any potential conflict of interest that includes any relationship between reviewer and author, or any content that is substantially similar to that under review.
- To conduct themselves fairly and impartially.
- To keep the confidentiality of the review process and to not retain or copy the manuscript.

Authors' responsibilities
- To ensure that their work submitted to the journal is original and authored by them and has not been previously published nor under consideration or accepted for publication elsewhere.
- To ensure that original ideas, data, findings and materials taken from other sources (including their own published writing) are properly documented and cited. Any content reproduced from other sources author should have permission.
- To ensure that their data is their won, true and not manipulated. Thus, authors are responsible to maintain accurate records and to provide access to their data associated with their manuscript.
- To ensure their work does not violate any rights of others, including privacy rights and intellectual property rights. In addition, authors should ensure that any studies involving human or animal subject are in accordance with local lows and requirements.
- To declare any real or apparent conflicting or competing interest at any stage during the publication process that could be considered or viewed as exerting an undue influence on his/her duties.
- To alert PEN in case if a significant error in their publication is identified and correct any errors prior or subsequent to publication of their work.
- To adhere to all research ethics guidelines of their discipline, and to ensure that authorship and/or co-authorship of the paper was accurately represented.

PEN responsibilities
- The journal and the publisher shall ensure that good practice is maintained to the standards outlined above.
- To deal with research misconduct allegations appropriately when it is occurred. The journal and the publisher shall undertake reasonable actions to identify and prevent the publication of papers where research misconduct has occurred.
- To maintain the editorial independence of journal editors.
- To support journal editors in ethical and academic matters.
- To ensure critical and objective assessment of all articles by reviewers and referees. To keep an accurate and transparent record, including publishing corrections and retractions when necessary.
- The following commonly recognized procedure for dealing with unethical behavior is adopted.

2. Procedures for Dealing With Unethical Behavior
Identification of unethical behavior
- Misconduct and unethical behavior may be identified and brought to the attention of the editor and publisher at any time, by anyone.
- Misconduct and unethical behavior may include, but need not be limited to, examples as outlined above.
- For an investigation sufficient information and evidence should be provided. All allegations should be taken seriously and treated in the same way, until a successful decision or conclusion is reached.

Investigation
- An initial decision should be taken by the editor, who should consult with or seek advice from the publisher, if appropriate.
- Evidence should be gathered, while avoiding spreading any allegations beyond those who need to know.

Minor breaches
- Minor misconduct might be dealt with without the need to consult more widely. In any event, the author should be given the opportunity to respond to any allegations.

Serious breaches

- Serious misconduct might require that the employers of the accused be notified. The editor, in consultation with the publisher should make the decision whether or not to involve the employers, either by examining the available evidence themselves or by further consultation with a limited number of experts.

Outcomes (in increasing order of severity; may be applied separately or in conjunction)

- Informing or educating the author or reviewer where there appears to be a misunderstanding or misapplication of acceptable standards.
- A more strongly worded letter to the author or reviewer covering the misconduct and as a warning to future behavior.
- Publication of a formal notice detailing the misconduct.
- Publication of an editorial detailing the misconduct.
- A formal letter to the head of the author's or reviewer's department or funding agency.
- Formal retraction or withdrawal of a publication from the journal, in conjunction with informing the head of the author or reviewer's department, Abstracting & Indexing services and the readership of the publication.
- Imposition of a formal embargo on contributions from an individual for a defined period.
- Reporting the case and outcome to a professional organization or higher authority for further investigation and action.

Originality and Plagiarism Policy

Authors by submitting their manuscript to PEN declare that their work is original and authored by them, their work submitted to PEN has not been previously published, original ideas, data, findings and materials taken from other sources (including their own published writing) are properly documented and cited, their work does not violate any rights of others, including privacy rights and intellectual property rights, their data is their won, true and not manipulated. Plagiarism in whole or in part without proper citation is not tolerated by the Journal. Manuscripts submitted to the journal will be checked for originality using anti-plagiarism software.

Comparative analysis of AMF, JSON and XML technologies for data transfer between the server and the client

Munir Šabanović	Muzafer Saračević	Emruš Azizović
University of "Džemal Bijedić"	Department of Computer sciences,	Faculty of Computer sciences
Faculty of Information Technologies	University of Novi Pazar	University of Prizren
Sjeverni Logor br.12 88000 Mostar, BiH	Dimitrija Tucovica bb, Novi Pazar, Serbia	Rruga e Shkronjave, nr. 1, Prizren, Kosovo
munir.sabanovic@uinp.edu.rs	muzafers@uninp.edu.rs	emrus.azizovic@unhz.eu

Abstract

This paper presents a comparative analysis of data formatting technology in AMF, JSON and XML, during data transfer between client and server. Data is authoring application that allows comparative analysis of these three technologies. These include aspects of the data transfer speed, the size of the output file, the data transmission safety, the code complexity of both the client and the server.

Keywords: AMF; JSON; XML; client-server communication; data transfer;

1. Introduction

The paper presents an authoring application, which allows measuring the time of data transfer from the moment of sending a request to the server, up to the moment of presenting data to the user. The measurement is realized for three data packaging technologies AMF, JSON and XML. The number of data from the base is chosen by the user in combo box, ranging from 1 to 76000. Time required to transfer data from the server to the client can be divided into several stages. Each phase is analyzed in detail and it is determined which of them introduces a delay during the data transfer. On the difference in time of data transfer, in addition to transfer technology, also affects the size of the output file, the complexity of the code of both the client and the server. The paper analyzes the aspect of data transmission security. At the end are given recommendations when and which technology to use.

2. Used scientific methods and procedures

In order to achieve the objectives and tasks of the research the following scientific methods and procedures were used:

- By comparative methods were compared results of the existing methods and new proposed methods of problems solving.

- By experimental application we have got results that demonstrate the effectiveness of new methods (in speed or saving memory space).

- The method of generalization is applied in the analysis of a number of cases where there is a general statement that applies to all the cases. The method of specialization presents specific cases as a specific example.

- Methods of deduction and induction are used in the course of experimental tests, where after the obtained results a conclusion is formed about the new techniques and the possibilities of their application.

3. Technologies used for making authoring application

For creating software solutions Flex technology was used on the client side, PHP on the server side, while, for data transmission were used three technologies AMF, JSON and XML. JSON and XML are used in many areas [6,7,11,12].

Flex is a highly productive, free open source software environment for creating executable version of expressive mobile, web and computer applications. Flex enables creating executable Web version and mobile applications that share a common basic code, thus shortening the time and cost of creating applications and long-term maintenance [1].

While Flex applications can only be created using the free Flex SDK, the Adobe Flash Builder ™ software can accelerate development with features such as intelligent code editing, gradual debugging, programs for memory optimization and performance as well as visual design.

Flex in its environment contains two MXML languages for visualization and AS3 for functionality [2].

Server PHP language, is executed on the server side. PHP is called a scripting language because it is written in the form of scripts and was purpose built for use on the Web [7]. PHP can be written in separate files, and can be inserted within HTML.

The PHP processor on the Web server interprets the PHP code, and the Web server on exit emits HTML or other types of data that the client web browser can understand. A copy of the HTML page is sent directly from the server to the client's computer, while the PHP code is not sent directly, but is before that translated into a form that the client computer will know how to interpret. HTML parts in the script are left as they are until PHP code interprets the PHP processor and executes, and the result of execution is sent to the client.

PHP code has great possibilities, of communicating with other computers, creating images, access to databases, work with graphics, creating desktop applications using PHP-GTK extensions all the way up to reading and writing files. PHP does not have its owner, it is a free language, a group of enthusiasts gathered together and made the PHP. For AMF technology, they took the Action Message Format and made serializators that correspond to action message format in PHP.

AMF or Action Message Format, is a binary format that is used for serialization of objects and sending messages between the client and the remote service. Action Script 3 language has classes for encoding and decoding of the AMF format. Adobe Systems has released AMF binary protocol specification on December 2007, and announced that it will support the developer community in making this protocol for all major server platforms. So today there is AMF support for platforms written in Java, PHP, .NET and other languages.

This paper will focus its attention on the PHP language, because it is the chosen server platform. JSON or JavaScript Object Notation is a very simple text format for data exchange between the server and the client [6, 10]. It is easy to parse and independent from any programming language. Parsing JSON format can be performed using the built-in JavaScript functions [11,12]. Compared to XML format, it is faster, shorter, there are no reserved words [8, 9].

XML, or Extensible Markup Language, is a language for creating electronic documents. One XML document is a hierarchy of XML elements. Each element represents part of the information contained in the document [3]. The content of the XML document includes: processing instructions, elements, attributes and comments.

4.Design and application functionality

Applications' operating environment contains a grid, with columns id, First name, Last name, department, index no. and Date. Figure 1. shows the applications' operating environment.
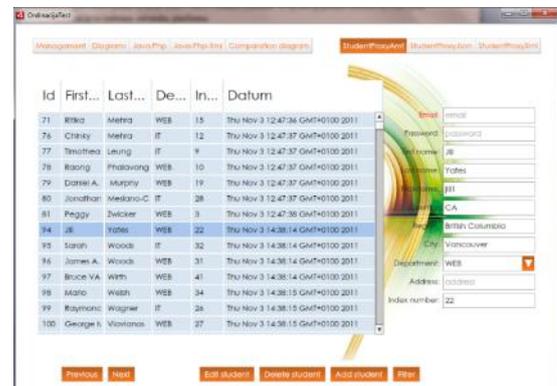
Figure 1. Applications' operating environment.

The columns are loaded data from the database, which is located on the server Wampserver2.4-x86. Number of loaded data is selected from the *ComboBox*, the range 1-76000 data.

Figure 2. DataGrid component on applications' workplace.

The application contains two Bar buttons. The first Bar button, *stateSelector* contains five states: *Managament, Diagrams, Java-PHP, Java-PHP-XML and Comparation* diagram. This control is defined by block code [3]:

```
<s:ButtonBar
top="25" left="30"
dataProvider="{statesProvider}"
labelField="label"
id="stateSelector"
change="stateSelector_changeHandler(event)"
/>
```

The dataProvider data supplier is a series of statesProvider, which is defined by the following block:

```
<s:ArrayList id="statesProvider" >
<fx:Object      label="Managament"      state=
"managament"/>
<fx:Object label="Diagrams" state="diagrams"/>
<fx:Object label="Java-Php" state="javaPhp"/>
<fx:Object      label="Java-Php-Xml"      state=
"javaPhpXml"/>
<fx:Object      label="Comparation      diagram"
state="comparation"/>
</s:ArrayList>
```

The authoring application loads the data from two different servers, Wampserver2.4-x86, which works with PHP and Niti server, which works with Java. If the active state is Managament, then the work environment of the application is displayed. However, if the second state Diagrams is active, then are displayed graphs for three technologies of data transmission, AMF, JSON and XML. The state JavaPhp shows JSON charts, when there is client authoring application communication with the Wampserver2.4-x86 server and Niti server. If the fourth state JavaPhpXml is active, then graphs are displayed which measure the time of n data, when the data on the part of the server is formatted in XML. If the Comparation diagram button is active, then a comparison chart of AMF-JSON-XML query execution is displayed.

The second Bar button selectProxyBar, changes its appearance depending on the application's state. The state Management, displays three buttons StudentProxyAmf, StudentProxyJsonand StudentProxyXml.

These three buttons offer choice of technology of data transfer from the database to the application in AMFU, JSON or XML.

Button controls, whose labels are Previous, Next, Edit student, Delete student, Add student and Filter allow, to be respectively displayed from the base the previous block of data in the grid, the next data block, edit the selected item in the grid and the database, delete rows in base, add new studentin the database, and filter data in the grid. Depending whether the state inclass instancesStudentControlComponent and StudentManagerComponent is normal or search, the filter control can take two values for label Filter and Back. TextInput fields and ComboBoxare used to show items from selected grid rows, changes to the content in the selected rows and adding new content.

In the state Normal, all the TextInput controls and ComboBox are shown, while in the state search are not displayed all TextInput fields.

5.Experimental results and comparative analysis

The application measures the time of data transfer from the server to the client for three technologies of data formatting, AMF, JSON and XML. Execution time requires measurement from the moment of sending a request to the server from client's part, until the time the data is downloaded from the server in the Flex application. JSON, XML, AMF are protocols used for communication between the client and server. Between them there is a difference in data packing speed, the manner of data formatting, the complexity of creating applications, the size of the file that is being transported,

as well as the complexity in terms of parsing data from one format to another.

AMF works with binary data, on the server's side it translates a language into binary data, while on the client's side the inverse process is performed, that translates zeros and ones into visual code that is displayed within the graphical interface.

In JSON technology, is carried out the translation of PHP objects into JSON objects, and eventually the JSON objects are sent as text file to the client through HTTP connections. The same is done with XML.

Working application uses AMFPHP package for the communication of client's ActionScript3 and server PHP language. In addition to the AMF, ActionScript3also supports XML and JSON technologies, performing data parsing to be transported in XML and JSON format.

During the data transfer between client and server, there are several phases, with each phase requiring certain time. On the client's side (Flex) there is a period for preparing the queries (serialization) and deserialization period, while on the server's side there is deserialzation period, the period query execution in the database, and serialization period. There is also a data transfer period. These periods are shown in Figure 3.
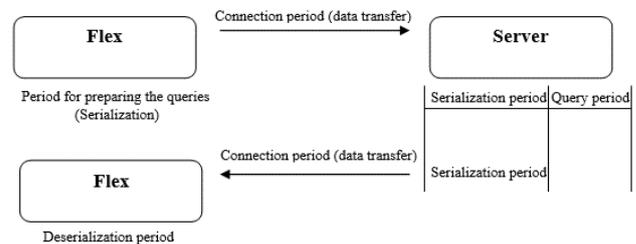
Figure 3. Period display, during data transfer from the server to the client.

Queries are extremely rapid for any technology, there are in microseconds and that period is not taken into account. Deserialization period on the server's side includes converting binary data into objects of some language. Query period is spent on reading data in the database, the data is then serialized, and then transferred to the client. And at the end, the data on the Flex side are deserialized for a certain time. Serialization has a global meaning, namely, in computer sciences, in the context of data storage, the serialization represents a process of translation data structure or objects in a format that can be saved (in the file or buffer memory, or can be sent through the network) and reconstructed, later in the same or different computer environment [5].

The production of AMF has, from the version 2.0, been undertaken by another company, and became considerably slower than JSON and XML. Slowness is due to the serialization, where a lot of time is spent on

converting objects of a language into binary data, which can be seen in Figure 4. This problem is solved by adding a plug-in, Baguette AMF, to the basic version of AMFPHP.
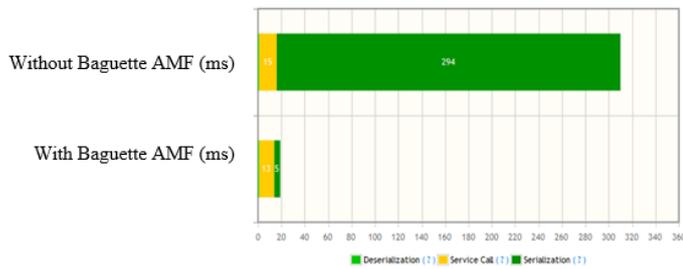
Figure 4. Period of serialization, deserialization and establishing a connection, when add-on Baguette AMF not in use and when it is used.

The application analyzes three important periods, the period of serialization, the period of deserialization on the client and server sides, the period of establishing a connection between the server and the client. The middle line defines the execution period of the service call.

The period of establishing a connection with the AMF technology is approximately the same with and without the Baguette AMF add-on. The deserialization periods are slightly different in both cases. Figure 4. shows that data serialization period drastically varies with and without Baguette AMF. The shorter serialization with Baguette AMF presents acceleration of the AMF.

Adding Baguette AMF plug-in to the basic version of AMF PHP, makes the AMF faster than JSON and XML when transporting larger amounts of data thus providing the Baguette AMF. Baguette AMF with AMFPHP is recommended for enterprise applications which share a large amount of data.

This paper translates PHP objects in AMF, JSON or XML format which is readable to the Flex client. Flex can serialize data in AMF, JSON or XML format and deserializate data from these formats. Figure 5. shows the communication between client and server where PHP is used as server language and serialization and deserialization of data is done with Baguette AMF.

The Client, within itself, has a built-in AMF format, which can be seen in Figure 5. AMF, XML and JSON technologies are widely used. They belong to cross technologies which means they do not depend on the operating system nor from the computer environment.
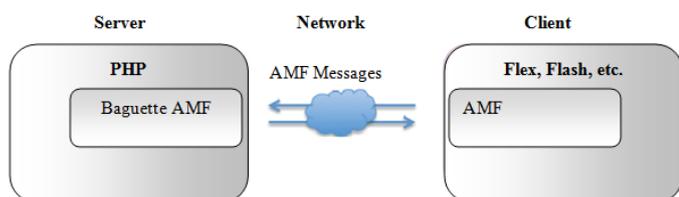
Figure 5. Flex – PHP communication

In the application, when the component Diagrams is active (see Figure 1), the charts for AMF, JSON and XML technology are shown. Comparison graph of data transport speed from the server to the client, which covers serialization period, deserialization, the period of establishing connection, is shown in Figure 6. In ComboBox is performed selection of data retrieved from the database on the server into the client application.
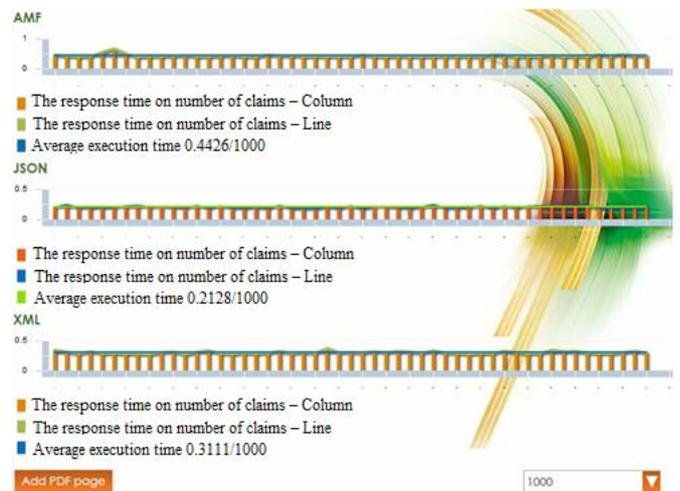
Figure 6. Graphs of time dependency from the number of loaded data for AMF, JSON and XML-technology

The choice of transfer technology is conducted by the selectProxyBar button bar. For each technology, fifty loading iterations are performed where in each iteration the same number of data is loaded, the time is measured, and at the end of the iteration the mean time is calculated.

Testing is performed on computer with the following performances: *CPU - Quad Core AMD Phenom X4 9550, 2200 MHz (11 x 200), L1 64 KB per core, 512KB L2 per core (On-Die, ECC, Full-Speed), 2 MB L3 (On-Die, ECC, NB-Speed), RAM Memory - 3 Gb, Graphic card - NVIDIA GeForce 9400 GT*. The environment of authoring application was used for the test.

On the graph each column presents the data loading period. Horizontal fault line connects the tops of columns and the right horizontal line in the chart describes the mean loading data period. The graphs represent time when 1000 data from the database are loaded.

On large number of data, AMF is faster than JSON. In Figure 7., on the basis of measurements in this model, is presented comparison diagram of dependency between the time and the number of data. Figure 7. shows, that up to 2000 data JSON is faster than AMF and above this number, AMF becomes faster. Therefore, it is recommended to use AMF technology above this number of loaded data. The diagram shows time dependency of the number of data downloaded, with JSON and XML it is much higher compared to the AMF. The Curve which describes the AMF technology, the third line, is of

considerably smaller angle, while with JSON, the first line is substantially steeper which means that the time is increasing faster with the number of loaded data. The middle line represents the XML format where we can see that the XML much more depends on JSON when the number of transmitted data increases.

Figure 7. - Comparison diagram of the data loading speed for AMF, JSON and XML technologies, for range from 100 to 2000 data and the step 100.

AMF does not work with HTTP connection. The advantage of AMF is opening of a socket connection, through which the client is forwarded the binary data. In Flex, in AMF technology, RemoteObject class is used to establish a connection to the server which inherits the Proxy class. AMF performs asynchronous communication, where data is sent and received without the agreed sending timing and receiving data.
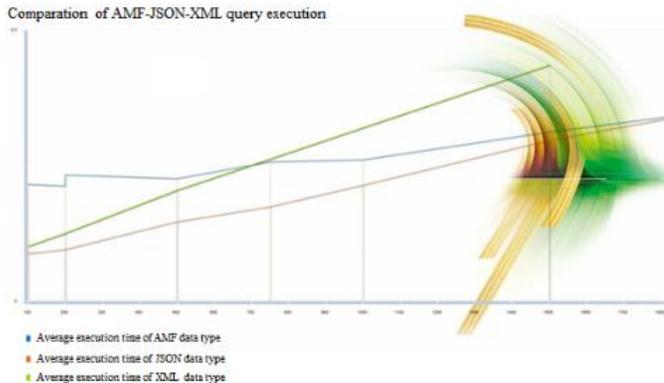
In AMF, the period of data transfer does not depend on the amount of data, because it is binary transmission and is always practically the same.

Figure 8. Structure of JSON file which is loaded into the client Flex application

In Figure 8. is displayed the contents of the JSON file when 10 data from the server are loaded in Flex application.

In JSON and XML the connection period depends on the amount of data, the data is packed in a file in text, and the time of the connection depends on the size of the file. Therefore, the more data, the connection period increases with JSON and XML except that the connection period with XML is bigger than with JSON for the same amount of data because of the structure of XML resulting in increasing the size of the output file and thus higher data transfer period.

Figure 9. displays the contents of an XML file when 10 data from the server are loaded to the client's application. Comparing the same content, in Figures 8 and 9, it can be concluded that XML file is more complex than JSON file, it has larger capacity and if both are legible.

Figure 9. XML file structure which is loaded into the client Flex application

Therefore, for the transfer of XML files, due to higher capacity, more time is needed, and due to the complexity of the XML structure, on the server and client side are used more complex classes to handle such structure in relation to JSON and AMF content all of which increases during serialization and deserialization.Therefore, if it bears the same data amount, because of the very structure of XML, the XML file is bigger than JSON. In Table 1. are given data on the size of the output file depending on the number of loaded data from the database for JSON and XML file. Data were measured while loading the application.

Based on the sample from Table 1., from 10 to 100 loaded data, the difference in file size between JSON and XML file is approximately 84%. Therefore, the XML file is 84% larger than the JSON file and thus the transfer of the same amount of data is shorter in JSON format.

N	10	20	30	40	50
A	2,73	5,41	8,01	10,44	13,05
B	5,03	9,97	14,75	19,36	24,19
C	84	84	84	85,3	85,4
N	60	70	80	90	100
A	15,7	18,36	21,16	23,97	26,43
B	29,03	33,88	38,89	43,9	48,54
C	84,9	84,5	83,8	83,1	83,6

A-Size of JSON file(Kb)
B-Size of XML file(Kb)
C- JSON file is bigger than XML file for (%)

On the side of the server and the client, the structure of XML data is more complex and such a structure is difficult to treat because it has more nodes. So, if there is a complex structure of the object, such as XML, it is more difficult to treat in any language because it has to go through more nodes, through greater depth. Therefore, when it comes to serialization and deserialization, there is no difference between languages. Each language will process faster the less complex data structure. In this model, in PHP, it is easier to process JSON objects from XML objects. PHP functions for parsing the XML object are more complex than when processing simpler object.

The data speed transfer is affected by the time serialization and deserialization, for any technology, the more data there is, the serialization and deserialization period is greater. Serialization is for the loop, which passes through a huge number of objects turning them into technology. For all three technologies the procedure is the same. For JSON there is inquiry into the base, and then the response from the database is converted into JSON and sent to the client. The client reads JSON, converts it into objects and continues to execute the command. The same applies to XML and AMF.

The serialization period on the server's side depends on the number of objects and if there are more objects to be serialized, then more time is spent which is especially emphasized in JSON and XML because the complexity is greater. Also the more data there is then the file in JSON and XML are heavier and more time is needed to transport data. Hence, the time difference is mostly reflected in the time required to transport data in AMF, JSON and XML and increases with the size of the file.

With 100 data in AMF, the average load time is 0,21s and with 1000 data it is 0,31s. In JSON, with 100 data it is 0,08s and with 1000 it is 0,30s, the difference being 3.62 times. This means that data transport has a major role. Various resources which are used in the application, can affect the choice of packaging data mode between the server and the client. For example, if a working application is using more services which use XML, and it is necessary for a service to be written, then it is logical to choose XML in order to have everywhere the same method of preparation and processing and thus to use the same class in the entire application for faster code writing and code consistency. Therefore, if one is using multiple data sources then it is up on the developer to create a single source. If they all use single technology, such as XML, then it is the developer's decision to choose the language in the part he deals with and that is XML because it will thus adapt the service made with other services in order to use class parts from other services and enable to assemble the two objects. Since if one object is in XML and the other in JSON, then the two objects cannot be assembled. In this case, both objects would have to be converted into a readable part in order to connect, which would slow data load. But if both objects are in XML or JSON, then they can be assembled because they are naturally compatible, there are methods for them, and that is one of the advantages that can be used.

When it comes to data security, SSL or encrypted communication mode can be used. The HTTP protocol is not safe when technology transfer JSON, XML or web services are used. The HTTP protocol can be used where data security is not that important because the content of JSON and XML file is readable (Figure 8.). SSL is a protocol for encrypting traffic. Here the traffic between points A and B encrypts so the content that is sent is not seen. Another manner is authorization of the services which can be made or bought and used. Authorization is function specifying the access rights to some resources associated with data protection. For authorization, the standard OAuth can be used which belongs to the open standard for authorization. OAuth provides the client application with a secure access to the resources allocated to the server on behalf of the resources' owner. This means, the client sends the server a request whether he could see his data.

The server responds so that it gives an affirmative answer, in terms of number or security code which the user will use through each subsequent request. That number in each query addresses the server as the data owner and that the server has allowed access to the client data. And each time the client receives a different password when he logs in. In general, the client is registered on the server as someone who wants to access server data. The server usually gives the client three data. The first data is the ID number of the application.

Therefore, the authoring application is registered on the server, as the owner of the data, and requires from the user that the application accesses specific data and then the server assigns ID to the authoring application when in the database the application ID is written down and as

the owner of the application the client is issued two data, a username and password. Now the client has three data, one is the application ID, the second is a username and password is the third. These three data are not used to access data, but another service is called for, which need not be on the same server where the data is.

The client calls for the service, passing the three data whereby following data verification, whether the application ID for that username is correct and whether the code is correct. If it is, the service generates one random number which is returned to the client in response authentication. At the same time in the database with each data is entered that same client's authorization number or it is placed in an authorization table. Figure 10. shows the schematic display of the authorization.
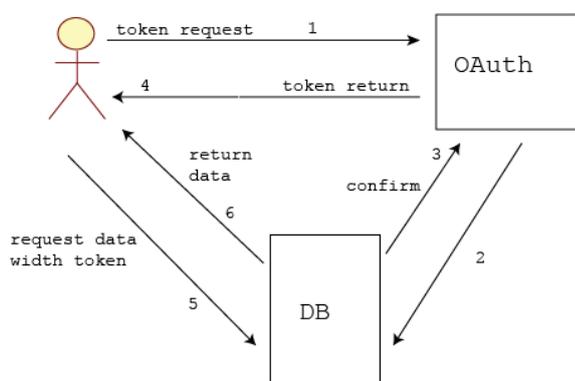
Figure 10. Access to the database with authorization

Following line 1, the user requires a key to the OAuth system. This system generates the key and following line 2 it is entered into the database DB. The DB basis following line 3, gives affirmative response that key was added. Following line 4, the key is sent to the client, that forwards it to the DB basis following line 5 while data requirement. Following line 6, the data is delivered to the client.

It is best to use both methods for data protection, authentication and encoding, with the SSL service. There are two ways of authorization: a time-limited and indefinite authorization manner. With time-limited authorization, the authorization service, issued the client permission to access data on the server for a limited time. After the timeout, the license to the client is terminated.

Either key is given to access the data for a period of time, which is calculated from the last access to the data on the server. This is typically done and on the web it is called session. This means, while the Client is active it gets unlimited time key. The time was measured from data requests, when a new request is sent it resets the time. When the time period expires, from the last request the system deletes the key from the authorization table. Facebook uses these things.

The things for authorization are linked to the http request, for JSON, XML, and Web services, since Web services are practically XML. The first call towards the server must be OAuth, the request for authorization.

6. Conclusions and Future Work

The choice of technology depends on several factors, one of them is the amount of data. By measuring data transfer period in the authoring application it is determined, for a small amount of data, up to 2000, JSON technology is recommended because it is faster, and it is closest to the object structure, orange variant on the graphics. When large amounts of data are used, up to 2000, AMF technology is recommended, blue variant on the graphics.

If using multiple sources that have been created in one technology, in XML, then it is recommended to the developer when creating the service to use the same technology, for the sake of simplicity, speed and consistency of code writing. When it comes to the complexity of the code, AMF technology is recommended, JSON and XML in the end. If it comes to the size of the output file, based on the measurements, JSON file is smaller for 84% of the XML file, and JSON file is recommended. If there is transfer of large data amounts, but segmentation is done, per segment is not transmitted more than 2000 data and then the recommendation is to use JSON technologies.

If it is a heterogeneous structure, which has support for XML, then it is advisable to use an XML format for communication between software and hardware components that communicate with each other. If data security is not important, then you can use the HTTP protocol for JSON, XML and Web services. If a more secure way of data transfer is important, then HTTPS connection is recommended.

7. References

[1] Using ADOBE FLEX 4.6. "Tutorial from official web site" help.adobe.com/ (visited 19.11.2014.).

[2] Zimmermann O., Tomlinson M., Peuser S. "Perspectives on web services: applying SOAP, WSDL and UDDI to real-world", *Springer*, 2003.

[3] Hall C., "ActionScript Developer's Guide toPureMVC", *O'Reilly Media*, 2011.

[4] C++ FAQ. "Kept by Marshall Cline", retrieved from www.parashift.com/c++-faq-lite/ (visited 19.11.2014.)

[5] Encryption Basics. "EFF Surveillance Self-Defense Project, Surveillance Self-Defense Project, n.d. Web". ssd.eff.org/tech/encryption (visited 19.11.2014.)

[6] Florescu D, Fourny G. "JSONiq: The History of a Query Language", *IEEE internet computing*, vol. 17, no. 5, pp. 86-90, 2013.

[7] Pettit, Jb, Marioni, Jc. "BioWeb3D: an online webGL 3D data visualisation tool", *BMC bioinformatics*, vol. 14, no. 1, pp. 185, 2013.

[8] Jorstad, I, Bakken, E, Johansen, Ta, "Performance evaluation of JSON and XML for data exchange in mobile services", *WINSYS 2008: proceedings of the international conference on wireless information networks and systems*, pp. 237-240, 2008.

[9] Rodrigues, C, Afonso, J, Tome, P, "Mobile Application Webservice Performance Analysis: Restful Services with JSON and XML, enterprise information systems", *Communications in Computer and Information Science*, vol. 220, no. 1, pp. 162-169, 2011.

[10] Adamanskiy A, Denisov A. "Embedded JSON database engine", *Fourth world congress on software engineering (WCSE)*, pp. 161-164, 2013.

[11] Merelo-Guervos J, Castillo J, Laredo A. et al.. "Asynchronous Distributed Genetic Algorithms with Javascript and JSON", *Conference: IEEE Congress on Evolutionary Computation Location: Hong Kong*, pp. 1372-1379, 2008.

[12] Jun Y, Zhishu L, Yanyan M. JSON Based "Decentralized SSO Security Architecture in E-Commerce, International Symposium on Electronic Commerce and Security", *Proceedings of the international symposium on electronic commerce and security*, pp. 471-475, 2008.

[13] http://www.tricedesigns.com/2011/11/07/AMF-vs-JSON-in-air-mobile-applications/ (visited 25.11.2014)

[14] http://www.jamesward.com/2007/04/30/ajax-and-flex-data-loading-benchmarks/ (visited 25.11.2014)

Modeling of Distributed Generators in 13 Nodes IEEE Test Feeder

Henry Giovanni Pinilla

Universidad Central

Bogotá, Colombia

hgpinillar@ucentral.edu.co

Andrés Julián Aristizábal

Universidad de Bogota Jorge Tadeo Lozano

Bogotá, Colombia

andresj.aristizabalc@utadeo.edu.co

Carlos Andrés Forero

Universidad de Bogota Jorge Tadeo Lozano

Bogotá, Colombia

carlosa.foreron@utadeo.edu.co

Abstract

This work aims to develop a model capable of evaluating the behavior of distributed energy resources in 13-nodes IEEE systems as a result of the change in the disconnector's opening protocol that creates a power generation island. The first scenario simulated a failure in the 632-671 line isolating the subsystem into two 375 kVA distributed generation units (DG) in the nodes 675 and 652. Likewise, a second scenario considered the aperture of the disconnector located between nodes 671 and 692 representing a 375 kVA DG feeding a 900 kVA load. The last scenario produced a three-phase failure modeling two 500 kVA DG units in the nodes 634 and 646 supplying an 800 kVA load.

Keywords: Distributed Sources, Electric Simulation, Generation Model, Islanding Operation, Power Systems.

1. Introduction

The distributed generation is defined as the utilization of electric power external sources directly connected to an existing power distribution infrastructure. These sources are denoted as Distributed Generation (DG) [1]

The distributed generation is based on the production of electricity located near the demanding load and frequently installed in the same building. It considers a wide variety of technologies depending on the availability. Sometimes, this distribution is named "disruptive" due to its propensity to affect the electric networks in industries. Mostly, microturbines and fuel cells will be the dominant components of distributed generation systems that will affect the dynamic grid components. [2]

Recently, there are two main pullers of distributed generation systems. The first puller is the increasing interest in the utilization of sustainable and clean energy resources. Likewise, the second puller is the new direction of the policies defined by some governments which allow independent companies to sell electricity using existing transmission and distribution grids. [3] Therefore, small-scale generation industries could find new opportunities to participate in the local energy market.

Large-scale installations guarantee the regional supply; therefore, they are not considered as a DG unit. These units are independent small-scale systems connected to the infrastructure due to economic and logistics reasons. [2,4]. The people is having a raising interest in installing power plants capable of supplying their own demand and feeding the exceeds to the grid generating extra income. Consequently, governments could decrease the high investments in the power generation sector resulting in a lower energy price and a high-quality supply [5].

The classical application of symmetrical components to short circuit computation is based on the premise that the normal pre-fault network is symmetricaland with balanced loads; while this assumption is acceptableat the transmission level, it does not hold in many distribution system feeders where the number of phases is less thanthree [6].

Distribution networkanalysis can be carried out either using phase coordinates orsymmetrical components, with the choice mainly dependingon whether calculations in one domain lead to computationalperformance improvements[7].

Existen en la literatura diferentes estudios relacionados con el modelado de sistemas de distribución teniendo en cuenta diferentes métodos, configuraciones y elementos que han contribuido de manera significativa a la investigación de redes de distribución de energía eléctrica a nivel mundial [8 - 12].

This work aims to develop a model capable of evaluating the behavior of distributed energy resources in 13-nodes IEEE systems as a result of the change in the disconnector's opening protocol that creates a power generation island. The stability analysis exposed here allows identifying the performance of a radial distribution power system with different disturbances. Results are relevant in the planning and operation stages of a distributed generation system.

2. Material and methods

Typically, the structure of a distribution system is based on radial or tree like feeders, probably with loops, which are often operated as open rings. In these networks, measurements are mainly located at the feeder-heads, either as voltage measurements, active and reactive power measurements or as current measurements [13].

Unlike transmission and subtransmission systems, where real-time telemetry provides sufficient redundancy to assurenetwork observability, medium voltage (MV) distribution feedershave so far lacked the required infrastructure (sensorsand telecommunication) allowing the operating point to beaccurately determined [14].

Tests were performed using synchronous and asynchronous machines. The synchronous machines were based on smooth rotor units of 1.25 MVA capacity [15, 16]. These units were characterized using the software NEPLAN [17]. The transitory and sub transitory impedance sequences were needed in the short-circuit modeling. The parameters were defined in accordance to [18] and are summarized in table 1 and table 2.

Modeling the central generation unit was done using parameters gathered in Table 3. The simulations were performed using a 13-nodes IEEE radial system [19].

A 615 MVA synchronous generator capable of supplying the load was employed. This generator replaced the grid feeder used in the IEEE test model. It was necessary to add a 5 MVA power transformer which decreases the voltage from 15 kV at the generator outlet to 4.16 kV; this is the voltage set in the system [15, 16] (Figure 1).

Table 1. Parameters Considered in the Synchronous Distributed Generation [18]

0.4 – 1.25 MVA 0.48 kV	T'do = 5,51s	D = 0	X''d = ,0171
	T''do = 0,10s	Xd = 2,062	X''q = 0,171
	T'qo = 0,8s	Xq = 1,35	Xl = 0,1
	T''qo = 0,1s	X'd = 0,251	S(1,0) = 0,176
	H = 1,29s	X'q = 0,631	S(1,2) = 0,49

Table 2. Parameters of the Synchronous Distributed Generation Exciters [18].

TR = 0,02s	VRMIN = -7,3	TF = 1,0
KA = 400	KE = 1	E1 = 5,475
TA = 0,02s	TE = 0,253s	SE(E1) = 0,5
VRMAX = 7,3	KF = 0,03	E2 = 7,3 SE(E1) = 0,86

Table 3. Parameters considered in the central generation system [18]

	T'do = 3,3s	D = 2	X''d = 0,23
615MVA 15kV	T''do = 0,02s	Xd = 0,898	X''q = 0,2847
	T'qo = 0,001s	Xq = 0,646	Xl = 0,2396
	T''qo = 0,06s	X'd = 0,2995	S(1,0) = 0,18
	H = 5,145s	X'q = 0,646	S(1,2) = 0,33

The aerial distribution lines were set up according to the characteristic of impedance and distance between the phase conductors [20]. The dimensions of each line varied for each segment depending on the nodes location.
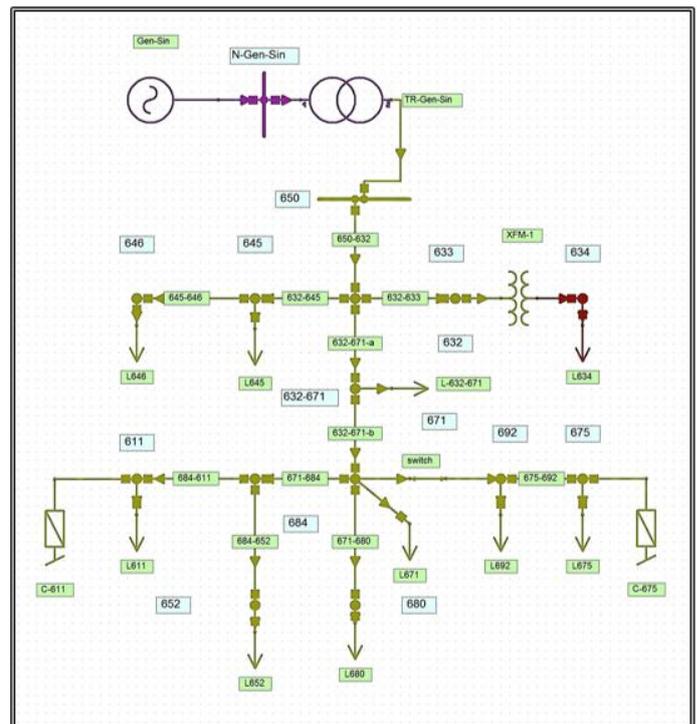
Figure 1. Layout of the 13-nodes IEEE radial test system [19]

Table 4 summarizes the impedance values and other configuration parameters.

Setup	R (ohm/mile)	X (ohm/mile)	B (µS/mile)	Ir max (A)
601	0,3465	1,0179	6,2928	730
602	0,7526	1,1814	5,6990	340
603	1,3294	1,3471	4,7097	230
604	1,3228	1,3569	4,6658	230
605	1,3292	1,3475	4,5193	230
606 (sub)	0,7982	0,4463	96,8897	329
607 (sub)	1,3425	0,5154	88,9912	310

The excitation system of the synchronous machine was modeled using NEPLAN; a control block was fitted and the recommendations mentioned in [21] were followed as shown in Figure 2.
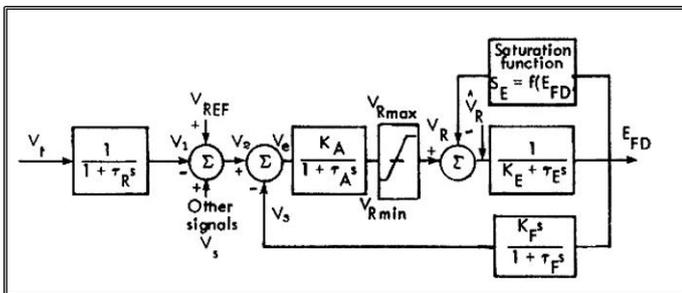
Figure 2. Type 1 excitation system [22]

The regulator of the excitation system works continuously in the model [22]. The control block and the synchronous generator were fitted, modeling the distributed generation system as shown in Figure 3.
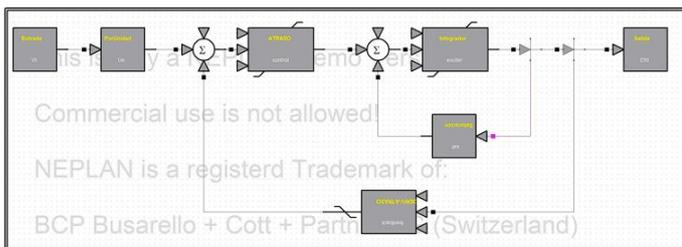
Figure 3. Control block applied to the type 1 exciter

The variables that were adjusted were included in Table 5. Parameters were defined in NEPLAN and the transference function was established [16].

Connecting the generators to the distribution grid was possible by using two transformer models. One is a 16 kV 15/4 reducer type that connects the central generation unit; whereas the second is a 16 kV 0.48/4 elevator type transformer which connects the synchronous generators to the nodes of the grid.

Table 5. Configuration of the Exciter[20]

Symbol	Description	Value
TR	Time constant of the regulator input	0.02 s
KA	Regulator gains	400
TA	Time constant of the regulator amplifier	0.02 s
KE	Gain exciter	1
TE	Time constant exciter	0.253 s
KF	Loop gain stabilization of the output voltage of the amplifier	0.03
TF	Time constant of the stabilization block	1.0 s
VRMAX	Maximum regulator output voltage	7.3 V
VRMIN	Minimum regulator output voltage	-7.3 V
E1	Saturation Voltage	5.475 V
SE75max	Exciter saturation function at 75 %	0.5
E2	Saturation Voltage	7.3 V
SEMAX	Exciter saturation function at 100 %	0.86

Parameters were included in Table 6.

Table 6. Transformers Configuration [20]

	S (KVA)	VH (kV)	VL (kV)	R %	X %
TR Gen-Sin	5000	15 – D	4,16 – Y	1	8
TR GD	400	0,48 – D	4,16 – Y	1,1	5
XFM-1	500	4,16 – Gr W	0,48 – Gr W	1,1	2

3. Results and discussion

3.1. Load flow analysis

The analysis of the steady state was done utilizing a load flow and the application of the Newton-Raphson method on the 13-node model. Based on this, we determined which nodes showed the largest drop on its voltage profile as shown in Table 7.

These are the most sensible nodes; therefore, they were considered as fundamental criteria to include the synchronous and asynchronous GD units together with the static and dynamic analysis [16]

The simulation of the system considering the central generation unit exclusively (Figure 1) exhibits that nodes 646 and 680 have the largest decrease in the voltage profile with values about 94.09 and 93.91%, respectively. These results are under the operational limit between 90 and 110%.

Table 7. Results of the Load Flow of the 13-node IEEE System.

MS 13 N								
Node	U kV	U %	Node	U kV	U %	Node	U kV	U %
646	3,91	94,09	684	3,938	94,6	675	3,928	94,42
645	3,92	94,26	611	3,934	94,5	N-Gen-Sin	16,5	110
633	4,11	98,99	652	3,929	94,4	650	4,421	106,29
671	3,94	94,83	680	3,903	93,8	N 632-671	4,034	96,97
634	0,45	95,03	692	3,943	94,7	632	4,134	99,37

3.2. Voltage stability

The voltage stability was studied doing several simulations of continuous load flow; these were performed varying the operational control of the DG unit in the PV or PQ mode, and changing the load factor when the asynchronous DG unit was added.

Figures 4 and 5 exhibit the effects that these changes have on the voltage stability observable in the voltage profile of the node 632. Results provide insights about the voltage collapse when the power transferences increases in a specific region of the radial system; nonetheless, the collapse point for all the nodes occurs at the same power level regardless of the voltage observable in the specific nodes.
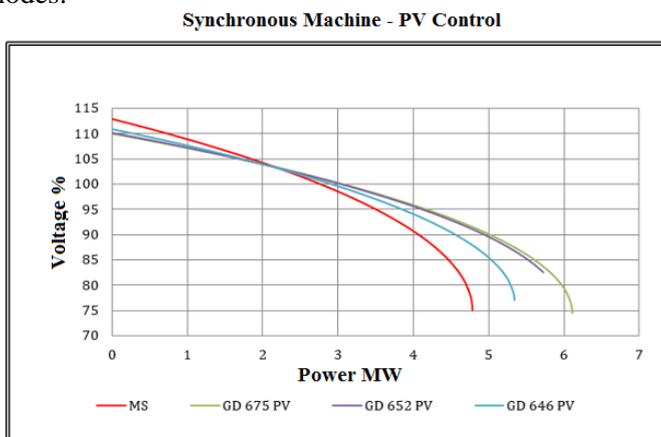
Figure 4. PV control on node 632

As a result of installing a generation unit in the node 675, 653 and 646, the stability of the systems and the load factor "λ" will rise in comparison to the base case where the load flow was done using the central generation unit solely (Figure 4). After adding the DG unit at the node 675, the chargeability moves up to 6.076 MW with a voltage collapse level of 77.305 V% .

Figure 5 shows the performance of the system when controlling the operation of the PQ machine.
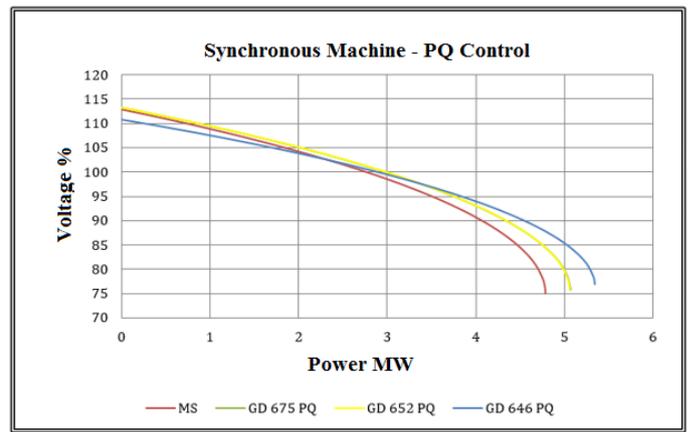
Figure 5. PQ control on node 632

The load factor in the node 646 and the voltage collapse level augmented up to 0.545 MW and 78.419 V% in comparison with the MS. This means that the load factor does not depend on the time; likewise, as a result of the change in this factor, the equilibrium points, and the system trajectories vary.

Simulations of the asynchronous machine (figure 6) exhibit that the load factor of the DG unit is lower than the 4.772 MW resulting from the central distribution system.
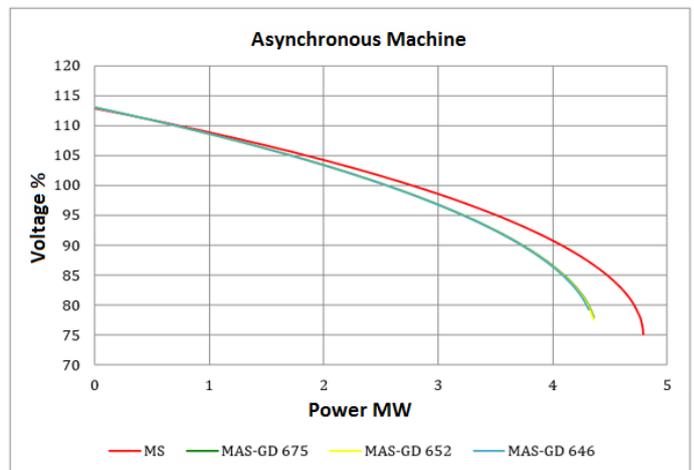
Figure 6. Asynchronous machine, node 632

Meanwhile, the voltage collapse limit is 77.093V% which reduces the stability margin of the system. This can be ascribed to the fact that asynchronous machines are not capable of generating reactive power.

It is noted that the profiles voltage increases in the most critical nodes.This indicates that the stability for the load increases when machines of generation are installed close or where larger loads are.Voltage profiles at 680 and 646 nodes increased approximately 3%.It is also noted that the voltage profile varies around 1.5% and depending on the

point of location on the network; profile increases by approximately 2%.This is very positive because if a disturbance occurs in the system, it is capable of returning in an acceptable time to a steady state where the voltages in all nodes are within an operating range.

3.3. Islanding operation analysis
Islanding operation conditions were adjusted by adding generation units in the nodes 675 – 652 – 646 – 634.

Afterward, three scenarios were simulated to analyze possible real failures in grids with distributed generation.

3.3.1. Failure in a line
The first scenario assumed a failure in the line 632-671 isolating the subsystem into two 375 kVA DG units in the nodes 675 and 652. These units fed a 1.7 MVA load as exhibited in figure 7.
A three-phase failure was promoted in the line in two seconds, and then an answer was observed in the switch located at node 671 which solve the failure in three seconds.
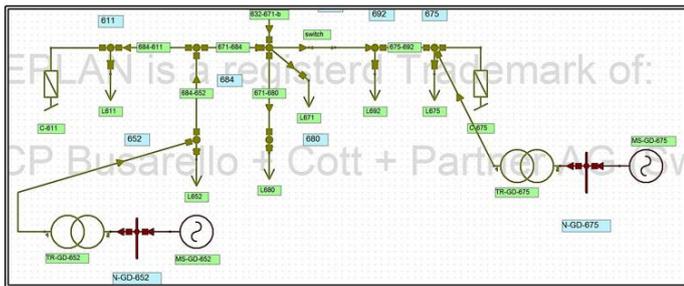
Figure 7. Layout of the system under islanding operation

Figure 8 shows the effect of the failure on the line. There is an oscillation that is supported by the central generation unit together with the two DG units and their corresponding excitation systems; nevertheless, the frequency gets out of synchrony when the switch placed at node 671 is opened.
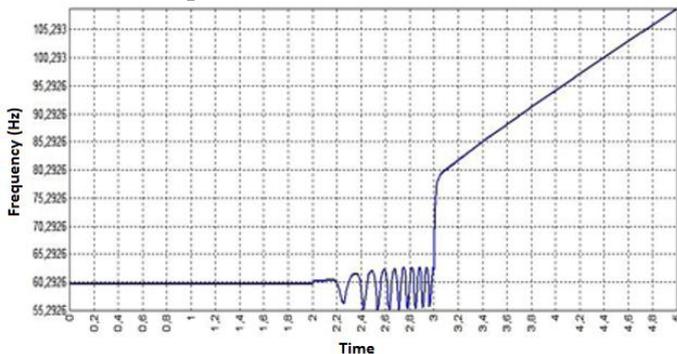
Figure 8. Performance of the frequency

Meanwhile, the DG units raise their frequency out of control because the central generation unit was supported the operational frequency of the radial network initially.

The active power provided by each unit undergoes some changes due to the failure simulated (Figure 9).
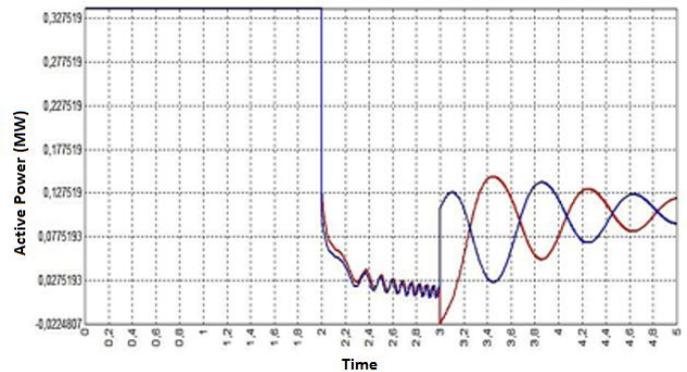
Figure 9. Performance of the active power

Exactly 2s of the simulation,active power experiences a fall from 0,33MW until 0,077MW; point which has a reaction to 3s and then unexpectedly increases to a peak value of 0,13MW. These results of active power, can be used to resize correctly, electrical networks that work with distributed generation; as the active power flow in both directions, grid becomes dynamic and this affects the size of the conductorsand protection systems.

As a consequence of this disruption, there is a tension collapse as shown in the node 671 (Figure 10).
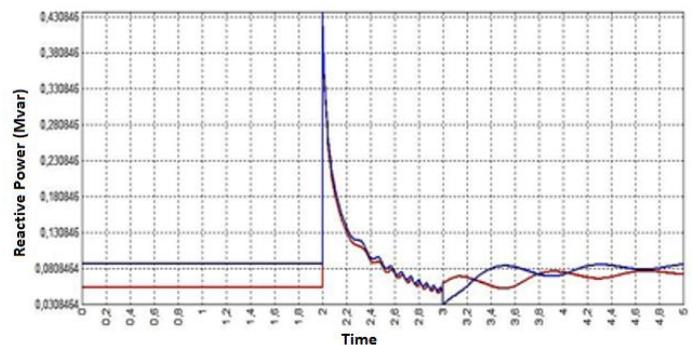
Figure 10. Results of the reactive power in the node 671

Furthermore, there is a less sensitive response when the DG units are asynchronous machines. Figure 11 exhibits the frequency oscillation when the failure occurs.
This parameter is recovered increasing the stability of the DG units after disconnecting the line 671.
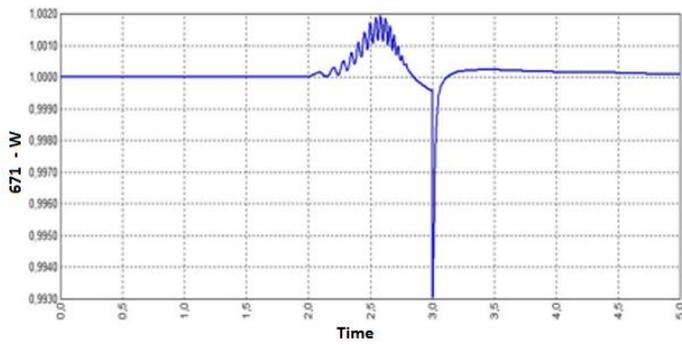
Figure 11. Frequency oscillation registered

These simulations show that the use of distributed generation sources increases the oscillation frequency of the 13 nodes system and these frequency oscillations are larger when asynchronous machines around 5.09 Hz are used.

Nonetheless, the active power drops to zero consequently with the voltage collapse as shown in Figure 12.
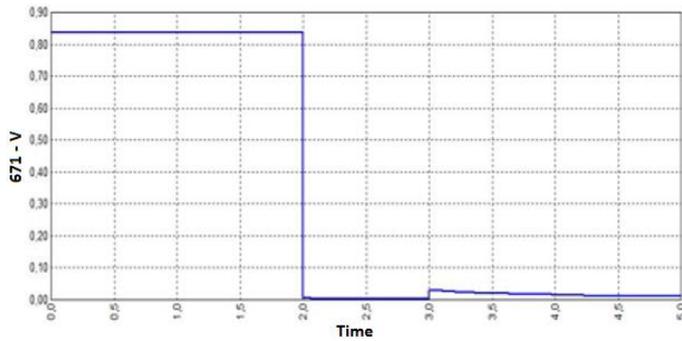
Figure 12. Voltage collapse registered in the node 671

This is ascribed to the high load demanded by the system in comparison with the power produced by the DG units. Finally, the reactive power, which DG units cannot supply, is satisfied by the shunt capacitance connected to nodes 611 and 675 (Figure 13).
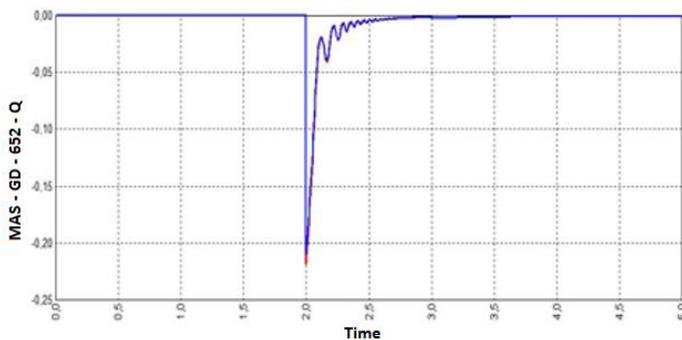
Figure 13. Effect of the shunt capacitance connected to nodes 611 and 675

3.3.2. Open a disconnector
The second scenario is simulated by opening the disconnector located between node 671 and 692. As a result of it, there is a 375 kVA DG unit feeding a 900 kVA load (Figure 14).
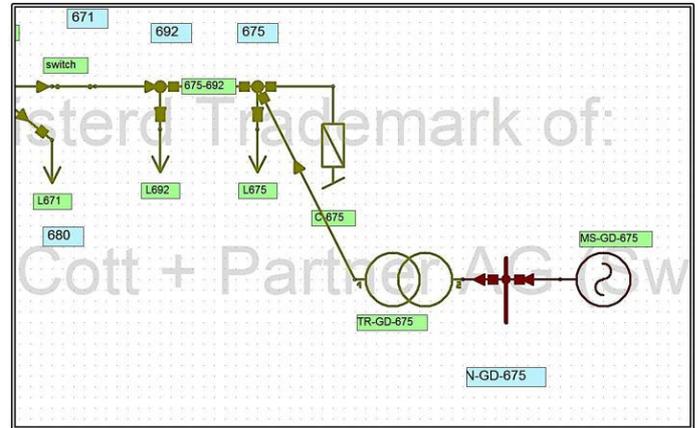
Figure 14. Layout of the system under conditions described in the scenario II

Disconnector is opened at 2s. Similarly as the scenario mentioned above, the disconnector aperture affects the synchrony of the system as shown in Figure 15.
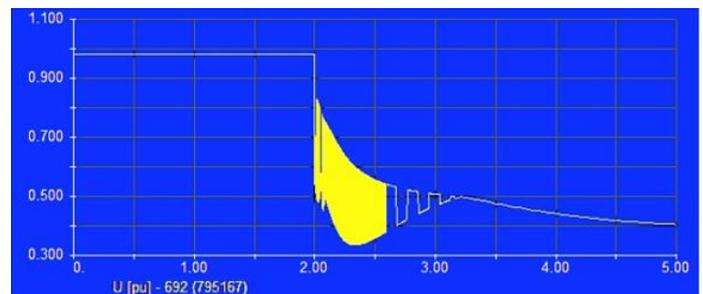
Figure 15. Voltage at node 692

The differences between the demand and the power capacity of the DG units produce a voltage collapse.

3.3.3. A three-phase failure
The third scenario consists of a three-phase failure occurring in the line 632 – 671 at 2s. The impedance and distance relays located in node 632 sense the failure, opening the disconnector attached to line 761 0.1 s later. Afterward, the impedance relay disconnects the central generation unit. This scenario models a system composed of two DG units in the nodes 634 and 646 with a power capacity of 500 kVA each one capable of supplying an 800 kVA load as stated in Figure 16.
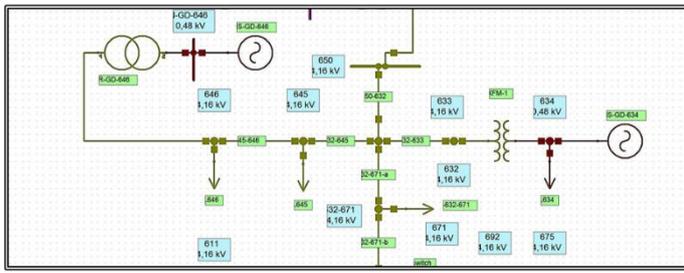
Figure 16. Layout of the system under conditions described in the scenario III

The performance of the two DG units is summarized in Figure 17.
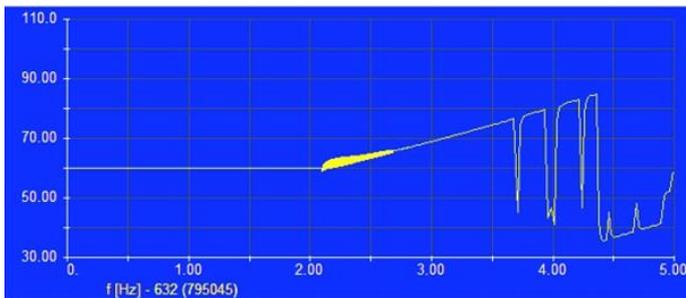
Figure 17. Frequency at node 632

The system frequency stabilizes after disconnecting the central generation unit; although, three seconds later, it varies again due to the load supplied by the two DG units. This load causes instability in the DG units, and a voltage collapse caused by the active power (Figure 18).
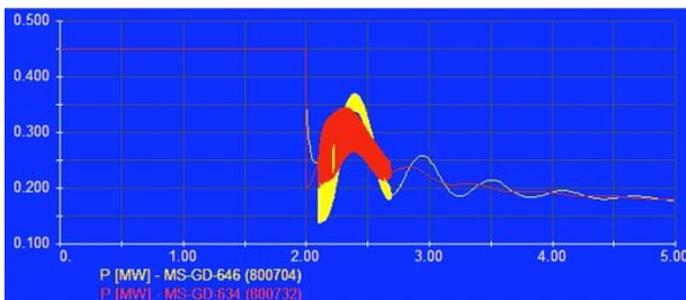
Figure 18. Active power in the GD units

5. References

[1] J. E. Kim, H. K. Tetsuo and Y. Nishikawa, "Methods of determining introduction limits of Dispersed Generation systems in a distribution system", ScriptaTechnica, Kyoto University, Japan, 1997.

[2] F. Jurado, J. Carpio. "Enhancing the distribution networks stability using distributed generation". The International Journal for Computation and Mathematics in Electrical and Electronic Engineering, Vol. 24, No 1, 2005.

[3] C. P. Lawrence, M. M. A. Salama and R. El Shatshat. "Studying the effects of distributed generation on voltage regulation", International Journal of Electrical Engineer Education, Vol. 46, issue 1, 2008.

4. Conclusions

The behavior of distributed energy resources in 13-nodes IEEE systems as a result of the change in the disconnectors´ opening protocol that creates a power generation island was modeled. The islanding operation of a power network consisting of distributed generation units have some security and stability problems that need to be considered. To keep the system operating conditions should be used controls and protections that detect failures and minimize their duration in the system, as recommended by the IEEE P1547.4 standard is a guide to the design, operation and integration of distributed generation systems in islanding power systems.

In cases where the system instability occurs and the island load is higher than the DG units capacity, protection measurements should be considered to sense the dynamic performance of the system and adjust the DG units or disconnect them when needed. Hence, voltage profile collapses can be avoided. Distance and impedance relays should be correctly parametrized because this can promote instabilities that will affect the dynamic response, cause large oscillations, and modify the power angle of the Dg units.

The adequate operation of the systems analyzed with DG units depends on the DG power capacity; it might be larger than the electrical load. The power balance should be preserved considering the synchronous or asynchronous characteristics of the machines. Also, AVR and PSS devices fitted to the generation units should be employed allowing to stabilize the system against fast and slow disturbances such as an aperture of switches, disconnectors, reconnectors, or changes in the load profile.

[4] R. C. Dugan and T. E. McDermott, "Operating conflicts for distributed generation on distribution systems", in Proc. IEEE Rural Electric Power Conference 2001, pp. A3/1–A3/6.

[5] N. Mithulananthanand Than O. "Distributed generatos placement to maximize the loadability of a distribution system", International Journal of Electrical Engineer Education, Vol. 43, issue 2, 2006.

[6] R. A. Jabr and I. Dzafic, "A Fortescue Approach for Real-Time Short Circuit Computation in Multiphase Distribution Networks," IEEE Trans. Power Syst., vol. 30, no. 6, pp. 3276–3285, Nov 2015.

[7] I. Dzafic, R. A. Jabr, and H. T. Neisius, "Transformer modeling for three-phase distribution network analysis," IEEE Trans. Power Syst., vol. 30, no. 5, pp. 2604–2611, Sept 2015.

[8] I. Dzafic, R. A. Jabr, E. Halilovic, and B. C. Pal, "A sensitivity approach to model local voltage controllers in distribution networks," IEEE Trans. Power Syst., vol. 29, no. 3, pp. 1419–1428, May 2014.

[9] I. Dzafic, B. C. Pal, M. Gilles, and S. Henselmeyer, "Generalized PI Fortescue equivalent admittance matrix approach to power flow solution," IEEE Trans. Power Syst., vol. 29, no. 1, pp. 193–202, Jan. 2014.

[10] I. Dzafic, M. Gilles, R. A. Jabr, B. C. Pal, and S. Henselmeyer, "Real time estimation of loads in radial and unsymmetrical three-phase distribution networks," IEEE Trans. Power Syst., vol. 28, no. 4, pp. 4839–4848, Nov. 2013.

[11] I. Dzafic, H. T. Neisius, M. Gilles, S. Henselmeyer, and V. Landerberger, "Three-phase power flow in distribution networks using Fortescue transformation," IEEE Trans. Power Syst., vol. 28, no. 2, pp. 1027–1034, May 2013.

[12] A. Gomez-Exposito, E. Romero-Ramos, and I. Dzafic, "Hybrid real–complex current injection-based load flow formulation", Electric Power Systems Research, vol. 119, pp. 237-246, Feb. 2015.

[13] I. Dzafic, I. Huseinagic, and S. Henselmeyer, "Real time distribution system state estimation based on interior point method," Southeast Europe Journal of Soft Computing, vol. 3, no. 1, pp. 32–38, Mar. 2014.

[14] A. Gomez-Exposito, C. Gomez-Quiles, and I. Dzafic, "State estimation in two time scales for smart distribution systems," IEEE Trans. Smart Grid, vol. 6, no. 1, pp. 421–430, Jan. 2015.

[15] Distribución Eléctrica Inteligente SILICE – Fase II, CODENSA, Universidad Nacional de Colombia, Universidad de los Andes, Colciencias, Septiembre de 2010.

[16] H. Pinilla, A.J. Aristizábal. "Análisis en estable y transitorio de fuentes de generación distribuida operando en isla", Revista Elementos, Numero 5, 2015.

[17] Software Neplan. [En línea]. Disponible: http://www.neplan.ch/neplanproduct/en-electricity/

[18] Abdel-Aziz A. Fouad Paul M. Anderson. Power System Control and Stability. IEEE Press, 2003.

[19] W. H. Kersting. Radial distribution test feeders. Technical report, IEEE Distribution System Analysis Subcommittee, 2000.

[20] IEEE Distribution Planning Working Group Report, "Radial distribution test feeders", IEEE Transactioins on Power Systems,, August 1991, Volume 6, Number 3, pp 975-985.

[21] IEEE Std 421.5-1992, IEEE Recommended Practice for Excitation System Models for Power System Stability Studies.

[22] Computer Representation of Excitation Systems, IEEE COMMITTEE REPORT, IEEE TRANSACTIONS ON POWER APPARATUS AND SYSTEMS, VOL. P.AS-87, No. 6 JUNE 1968.

Uncertainty Balance Principle

Migdat I. Hodžić
International University of Sarajevo, Bosnia and Herzegovina
mhodzic@ius.edu.ba, migdathodzic@gmail.com

Abstract

The objective of this paper is to present new and simple mathematical approach to deal with uncertainty transformation for fuzzy to random or random to fuzzy data. In particular we present a method to describe fuzzy (possibilistic) distribution in terms of a pair (or more) of related random (probabilistic) events, both fixed and variable. Our approach uses basic properties of both fuzzy and random distributions, and it assumes data is both possibilistic and probabilistic. We show that the data fuzziness can be viewed as a non uniqueness of related random events, and prove our Uncertainty Balance Principle. We also show how Zadeh's fuzzy-random Consistency Principle can be given precise mathematical meaning. Various types of fuzzy distributions are examined and several numerical examples presented.

Keywords*: Fuzzy Distributions; Cumulative Distributions; Fuzzy to Random Transformation; Random to Fuzzy Transformation; Uncertainty Balance Principle; Uncertainty Change Law; Big Data*

1. Introduction

In studies of uncertain phenomena, several methods are employed. Two most widely used are random and fuzzy data approaches. They are typically described in terms of random and fuzzy distributions [7]. These two methods look at the uncertainty from different points of view. In literature one can find various terms for fuzzy data, such as possibilistic, soft and subjective [3], as opposed to random called probabilistic, hard and objective [21]. These terms are somewhat arbitrary and there are authors who used probability distributions to represent subjective information [42, 43, 44]. Similarly, other authors used fuzzy sets and possibility distributions to describe objective imprecise information either about constants or about probability distributions [41]. Historically, probability is defined in the context of some physical measurement and mathematically in terms of probability axioms by Kolmogorov [34] where probability space, events and associated probabilities are defined. Related notion of random variables are defined in terms of mappings from probabilistic space of events to real line, carrying an underlying probabilities from the original event space. Indexing with some independent variable, such as time, one can define random processes as dynamic versions of random variables. On the other hand fuzzy, possibilistic approach relates to some intuitive uncertain notion (often of human nature) of an underlying uncertain event with some confidence (presumption) levels defined. Often in fuzzy data there is no reference, at least not directly, to any experiment or hard measurement. It is more representation of our confidence level in an uncertain phenomenon. If a need arises to combine fuzzy and random data, such as in soft/hard data fusion, [7],[21], each distribution is

typically handled separately for a specific problem at hand, and to our knowledge no rigorous mathematical methodologies exist for a practical uncertainty alignment between two types of data. In a fundamental paper by Zadeh [2], a concept of possibilistic fuzzy distributions was introduced as opposed to random and probabilistic distributions. The possibilistic distribution is shown to be equivalent numerically to fuzzy membership function. In classic fuzzy references [7],[8], various algebraic operations on fuzzy data are dscribed, as well as the methods as how to combine fuzzy and random data in meaningful ways. One obvious method is to normalize random data distribution to unity and combine it with the fuzzy data. Mathematically correct in principle, this method can be considered as a sort of uncertainty alignment from random to fuzzy data. Unfortunately the method is not practical because of loss of information in the process [7]. Also, in our opinion this method does not have any strong conceptual ground. Another approach is to define hybrid data which retains both fuzzy and random properties of original data. One can define random fuzzy data where fuzzy distribution argument is "randomized" according to a probabilistic distribution density. Or, one can consider fuzzy random data where the value of random distribution density is fuzzified according to fuzzy distribution. From these original ideas, there was very extensive development last two decades, [9]-[19] in the area of "random fuzzy sets" and "fuzzy random variables". Neither is the focus of our paper. The subject of our paper is to consider fuzzy to random uncertainty alignment (i.e. starting with fuzzy and generating random data, or vice versa) using very basic properties of fuzzy and random distributions. Our motivation is to produce a common

fuzzy or random data base to process the data further for either decision making process for a given application or a possible data filtering. In our approach we employ three step methodology, i.e.:

(I) Decompose any fuzzy distribution via cumulative (probabilistic) distribution functions (CDF). We do not use probabilistic density functions (PDF), a derivative of CDF, for two reasons. First, it may not always exist [34], and second is that CDF is normalized to unity by the definition, similar to fuzzy distributions.

(II) Use basic probabilistic axioms whereas the CDFs are defined in terms of random event probabilities of the form P(X≤x), [34],[35], and combined with (I) above resulting in probability differences $\Delta P(A_i)$ for some TBA events A_i.

(III) Use Big Data or some other statistical methodologies to produce best $\Delta P(A_i)$ choices in (II). a

The result of our approach is that for any unimodal fuzzy data, fuzzy distribution can be thought of as a combination of fixed and variable probability events. In the case of multimodal fuzzy data, this representation consists of a number of fixed and variable random events. We believe our approach can bring about new avenues in aligning fuzzy and random data, in particular in very important area of soft-hard (human-machine) data fusion [21]. In our previous introductory paper [40] we presented the basics of our uncertainty alignment methodology. This paper extends these results with additional fuzzy to random alignment methods, and it presents a unifying Uncertainty Balance Principle of the general form $\Pi(x) + \sum\Delta P(A_i,A_j) = 1$ and $\Pi(x) = \sum\Delta P(A_i,A_j)$ for all alignment cases. The paper is organized as follows. Section 2 summarize, very briefly, basic probabilistic and possibilistic results employed in this paper. We do not aim to be complete with these summary, just what is of interest for the current paper. In Section 4 we introduce our main fuzzy to random uncertainty alignment arguments using standard triangular fuzzy distribution (TFN). The approach applies to any fuzzy distribution. We describe steps which result in fuzzy distribution as a combination of fixed and variable probabilities. In Section 4 we formalize and prove key results:

(i) Probabilistic decomposition of an n-modal fuzzy distribution

(ii) Universal Uncertainty Balance Principle which is presumption and x-invariant, and related

(iii) Uncertainty Change Law.

In Section 4.7 we point to a potential use of our methodology in Data Fusion and Decision Making situations. Section 5 presents numerical examples showing fuzzy distributions in terms of fixed (unique) and variable (non unique) random events and related probabilities. Symmetric and non symmetric TFN, convex and non convex distributions are illustrated. The numerical examples confirm results of Section 4. Section 6 has the pseudo code for the uncertainty alignment algorithms, Conclusion is in Section 7, and key references are included in Section 8.

2. Random (Probabilistic) Distributions And Fuzzy (Possibilistic) Distributions

For the purposes of this paper we recall few basic classic probability and cumulative distribution facts as well as elementary fuzzy distributions results used in this paper.

2.1. Probability and Cumulative Distribution

The random events A, A_1, A_2, etc. are the subsets of a certain event S and they are assigned probabilities P. From classic references [34], [35] we have:

$$0 \leq P(A) \leq 1, \ P(S) = 1, P(O) = 0 \qquad (1)$$
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \qquad (2)$$

where O is an impossible event. If $A_1 \cup A_2 = S$ and two events are mutually exclusive, then $P(A_1 \cup A_2) = P(A_1)+P(A_2)=1$, hence A_1 and A_2 are complementary with $P(A_1) = 1 - P(A_2) = P(A_2^*)$ and A_2^* indicates complementary event to A_2. For any event A, the following holds:

$$P(A) + P(A^*) = 1 \qquad (3)$$

If the events are independent, then we have:

$$P(A_1 \cap A_2) = P(A_1)P(A_2) \qquad (4)$$

Mutual exclusivity and independency do not imply each other. A random variable $X(\xi)$ is a function that assigns a real number to each outcome ξ in the sample space S of a random experiment [34],[35]. If an event A is given in S such that A={ξ: $X(\xi)\epsilon B$}, where B is a subset of real line R, then A and B are equivalent events with the same probability:

$$P(X\epsilon B) = P(A) = P(\xi: X(\xi) \epsilon B) \qquad (5)$$

Cumulative distribution function (CDF) of X is defined as:

$$F_X(x) = F(x) = P(X \leq x), -\infty \leq x \leq +\infty \qquad (6)$$

which is a probability that X has a value in (-∞, x], and hence it is a function of x. Figure 1 shows uniform CDF and related PDF, which is a derivative of CDF. The properties of CDF and PDF can be found in any classic probability theory reference, such as [34] and [35].
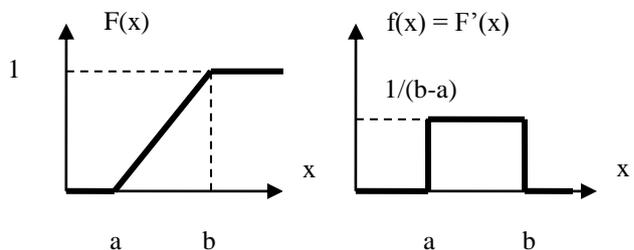
Figure 1. Uniform CDF and PDF

In this paper we deal with the cumulative rather than density functions (which may not exist in some cases), for mathematical as well as conceptual and practical reasons.

2.2. Possibility and Fuzzy Distributions

18

Possibility theory was developed early on by Zadeh [2]-[5] as an extension of fuzzy sets theory, in the context of information meaning, in particular in the context of semantic variables and human soft (fuzzy) data. Possibility was associated with fuzziness, either due to lack of knowledge or related to the subset for which possibility is defined. Since its inception possibility theory was developed in an axiomatic framework. We do not aim to discuss recent theoretical developments in possibility theory which there are many [9]-[19]. We simply recall possibility distribution $\Pi_X(\xi)$ as a fuzzy restriction on the values assigned to an uncertain variable X and numerically equivalent to fuzzy membership function $\mu_A(\xi)$, i.e. $\Pi_X(\xi) = \mu_A(\xi)$. For simplicity of the notation we use $\Pi(x)$, x representing specific choice of fuzzy variable X. Here we recall just a few fuzzy distribution properties which are used in this paper.

$$0 \leq \Pi(A) \leq 1, \quad \Pi(S) = 1, \Pi(O) = 0 \qquad (7)$$

where O is an empty set and S is a universe of discourse, with all subsets to which we can assign possibilities. For any A we also have:

$$\Pi(A) + \Pi(A^*) \geq 1 \qquad (8)$$

where A* indicates complementary event to A. The full axiomatic description can be found in [18].

2.3. Consistency Principle

We find it useful for our paper to recall Consistency Principle between fuzzy and random variable X defined in Zadeh's classic paper [2] as:

$$\Gamma_X = \sum P_i \Pi_i = P_1 \Pi_1 + P_2 \Pi_2 + \ldots + P_n \Pi_n \qquad (9)$$

where i=1,…,n and variable X can be interpreted both as probabilistic and possibilistic, with the corresponding distributions consisting of the same number of choices in the interval of interest. Consistency Principle carries an intuitive observation that reducing the possibility of an event tends to reduce its probability. The opposite may not hold. If there is a precise (point wise) match between possibilistic and probabilistic distributions, $P_i = \Pi_i$, then:

$$\Gamma_X = \sum \Pi_i^2 = \sum P_i^2 \qquad (10)$$

Consistency Principle as given in (9) may be useful when possibility is known about uncertain event X but not the probability. Our paper expands this idea via Uncertainty Balance Principle which produces variable probability from given possibility. In that sense our approach is similar to (10) rather than (9), as shown in Section 4. Our approach also lends itself to a precise mathematical and quantitative treatment.

3. *Fuzzy To Random Alignment*

Methodology described in this paper can be of good use in decision making process where data is inherently mixed, both soft (fuzzy) and hard (random). Often in literature one finds terms such as objective, sensor based, or machine for hard data, and subjective, human based for soft data. One particular area of interest is human-machine (soft-hard) data fusion [21]. The main contribution of our work can be understood as two fold:

(i) If fuzzy data is available, we can produce variable random data with variable probabilities reflecting original fuzziness.

(ii) On the other hand, if a variable random data is available we can produce corresponding fuzzy distribution, both contributions per our Uncertainty Balance Principle.

3.1. General Considerations

We proceed by considering a typical triangular fuzzy distribution number (TFN) given in Figure 2 with the interval of interest {a,b,c} for any a, b and c, and the corresponding fuzzy distribution $\Pi(x)$ numerically equivalent to the fuzzy membership function $\mu(x)$, [2]. At this point we will not write equations for the segments of $\Pi(x)$. This is done in Section 6 with numerical examples. Note that our approach can be applied to any other fuzzy variable, unimodal or multimodal, symmetric or not, normal or non normal, convex or non convex, trapezoidal or arbitrary shaped fuzzy distribution. Section 5.5 shows additional examples of fuzzy distributions. The key is that the CDF properties [35] are satisfied. Next step is to define a pair of CDF's such as in Figure 1 to "decompose" TFN distribution $\Pi(x)$:

$$\Pi(x) = F_1(x) - F_2(x) \qquad (11)$$

where $F_1(x)$ and $F_2(x)$ are shown in Figure 3. They are both uniform probabilistic distributions. The purpose of the decomposition (11) is a first step in relating fuzzy to random variables. Our first idea to define (11) came from an obvious fact that CDF is maximum at 1 similar to $\Pi(x)$. Next step is to find a way how to describe both rising and falling part of $\Pi(x)$, and hence (11) came as a natural solution. Recall that a CDF is a probability of an event A = "X ≤ x" as given in (6). It is critical we assume the uncertain variable X is <u>both possibilistic and probabilistic</u>.

Next we take another "probabilistic" step to refine (11) using basic probability relation in (2). The key is that equation (1) has negative term in it which we can associate with $F_2(x)$ in (11). This negative term can take different form depending how we move different terms around in (2). Three methods are described.

19

$\Pi(x) = \mu(x)$
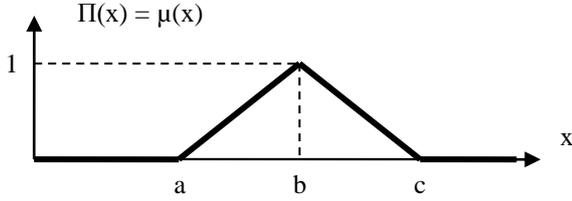
Figure 2. Triangular fuzzy number (TFN)

$F_1(x)$

$F_2(x)$
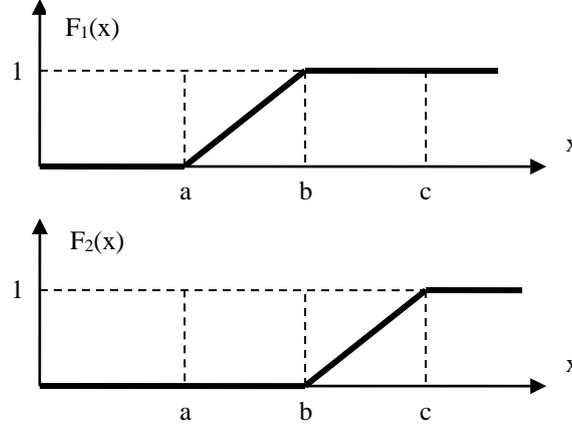
Figures 3. CDF's $F_1(x)$ and $F_2(x)$

3.2. Method 1

This method is described in full details in our earlier paper [40]. It is summarized here. We rewrite (2) as:

$$P(A_1) - P(A_1 \cap A_2) = P(A_1 \cup A_2) - P(A_2) \qquad (12)$$

Each side of Equation (12) is a probability and it satisfies basic probabilistic axioms in (1) and (2). Nice property of (12) is that both sides have negative terms, as does Equation (11). See also Figure 4. We proceed in two $\Pi(x)$ parts, rising and falling.

Rising Part. We equate left side of (12) with (11):

$$\Pi(x) = F_1(x) - F_2(x) = P(A_1) - P(A_1 \cap A_2) \qquad (13)$$

from where $F_1(x)$ and $F_2(x)$ could be uniquely associated with the corresponding probabilities in (13). The aim of this step is to formally align $F_1(x)$ with $P(A_1)$ and $F_2(x)$ with $P(A_1 \cap A_2)$, with an idea that an intersection of two events $A_1 \cap A_2$ can produce probability non uniqueness, whereas A_1 would be fixed. We again recall the fact that CDF is a probability of an event per (6). Next, the events A_2 and $A_1 \cap A_2$ (their probabilities $P(A_2)$ and $P(A_1 \cap A_2)$ given via $F_1(x)$ and $F_2(x)$) are to be determined. What is not uniquely determined is A_2 because different A_2 can produce the same intersection $A_1 \cap A_2$. A little reflection on set theory brings us to:

$$P(A_1 \cap A_2) \le P(A_2) \le 1 - \Pi(x) \qquad (14)$$

producing the same $P(A_1 \cap A_2)$. The right side $1 - \Pi(x) \le \Pi^*(x)$ represents complementary fuzzy distribution to $\Pi(x)$, which upholds the condition in (10). Note that non unique A_2 corresponds to $x \le b$, while A_2 is unique for $b \le x$, due to a simultaneous action of conditions in (14). One can consider that the interplay of unique $P(A_1)$ and non unique $P(A_2)$ in $P(A_1 \cap A_2)$, produces "fuzziness" on the left hand side of $\Pi(x)$. See also Figures 5.

Falling Part. We equate right side of (12) with (11):

$$\Pi(x) = F_1(x) - F_2(x) = P(A_1 \cup A_2) - P(A_2) \qquad (15)$$

where $P(A_2)$ is uniquely defined in (15). The events $A_1 \cup A_2$ and A_1 are to be determined.

$F_1(x) \quad P(A_1) \quad P(A_1 \cup A_2)$

$F_2(x) \quad P(A_1 \cap A_2) \quad P(A_2)$
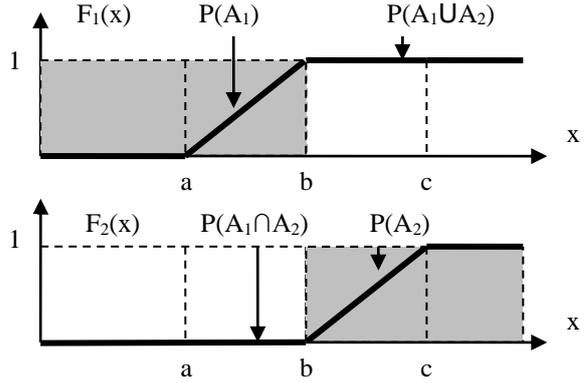
Figures 4. Method 1: CDF's $F_1(x)$ and $F_2(x)$

$1 - \Pi(x)$

$P(A_1 \cap A_2)$

$P(A_2)$
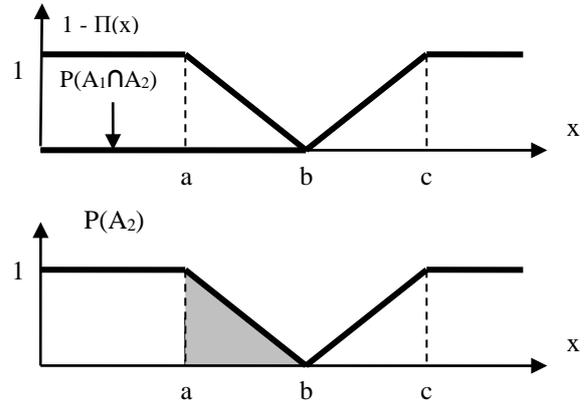
Figures 5. Rising part bounds on $P(A_2)$

$P(A_1 \cup A_2)$

$P(A_1)$
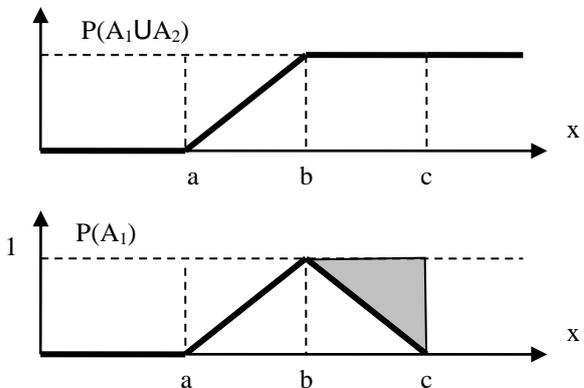
Figures 6. Falling part bounds on $P(A_1)$

Note that the forms of F_1 and F_2 are same as before but we interpret them differently, i.e. as $P(A_1 \cup A_2)$

20

and $P(A_2)$. To further clarify A_1, we consider union $A_1 \cup A_2$, which produces the following condition on $P(A_1)$:

$$\Pi(x) \leq P(A_1) \leq P(A_1 \cup A_2) \qquad (16)$$

Figures 6 summarize $P(A_1)$ and bounds given in (16) for $\Pi(x)$ using $P(A_1 \cup A_2)$. The gray area in $P(A_1)$ indicates its non unique choices. They will all generate the same $F_1 = P(A_1 \cup A_2)$. Note that the grey area corresponds to $b \leq x$, while $P(A_1)$ is uniquely defined for $x \leq b$, due to a simultaneous action of conditions (16). As in Rising Part, one can consider that the interplay of unique A_2 and $P(A_2)$ and non unique A_1 and $P(A_1 \cup A_2)$ is equivalent to "fuzziness" of the right hand side of $\Pi(x)$, when $b \leq x$. By combining two parts, we conclude that non unique choices for A_1 and A_2 and their corresponding probabilities $P(A_1)$ and $P(A_2)$, correspond to the non zero part of the distribution $\Pi(x)$. Outside of that, when $\Pi(x)=0$, they can be considered independent for the trivial cases of probabilities 0 or 1, per Table 1, where we used the notation $P_1=P(A_1)$ and $P_2=P(A_2)$. Note that the non zero $\Pi(x)$ corresponds to the gray shaded areas in Table 1.

Table 1. $\Pi(x)$, Method 1. Equations (16) and (19)

x	A_1	A_2	A_1 vs. A_2
$0 \leq x < a$	$P_1 = 0$	$P_2 = 0$	$P_1 + P_2 = 0$
$a \leq x < b$	$P_1 = \Pi$	Non unique	$P_1 + P_2 \leq 1$
$b \leq x < c$	Non unique	$P_2 = 1 - \Pi$	$P_1 + P_2 \leq 1$
$c \leq x$	$0 = P_1$	$P_2 = 1$	$P_1 + P_2 = 1$

Final note is that if we choose (13) for the Falling Part instead of (15) we end up with $P(A_1)$ and $P(A_2)$ as constant probabilities, given $\Pi(x)$, and there will be no "fuzziness" induced by variable probabilities.

3.3. Method 2

Now we use Equation (2) and define fuzzy distribution (11), with the following choices for CDF's $F_1(x)$ and $F_2(x)$:

$$F_1(x) = P(A_1) - P(A_1 \cap A_2) \qquad (17a)$$
$$F_2(x) = P(A_3) - P(A_3 \cap A_4) \qquad (17b)$$

Venn diagrams in Figures 7 are for $x \leq b$ with the arrows indicating two evens A_1 and A_2 "extending" to eventually form a certain event with the probability 1, with either $P(A_1 \cap A_2) = 0$ (A_1 and A_2 meet) or $P(A_1 \cap A_2) \neq 0$ (A_1 and A_2 overlap), with $P(A_1) + P(A_2) - P(A_1 \cap A_2) = 1$ in either case. This corresponds to the presumption (possibility) level of $\Pi(x) = 1$.
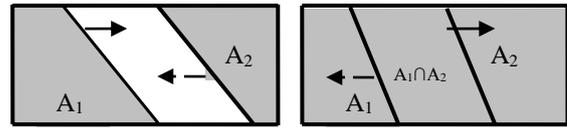
Figures 7. Formation of fuzzy distribution $\Pi(x)$ in (17a)
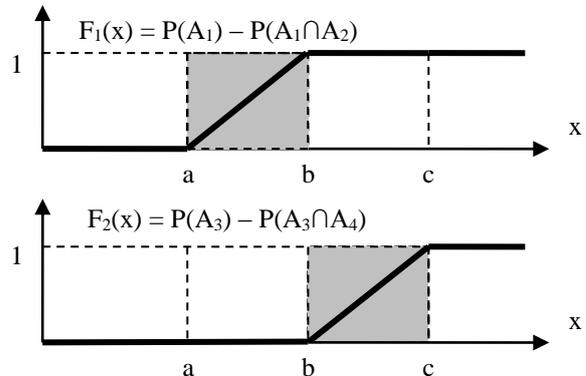
Figures 8 Method 2: CDF's $F_1(x)$ and $F_2(x)$

For $b \leq x$, the same process starts with two new events, A_3 and A_4 with the probabilities in (17b) for F_2. For $c \leq x$ we have $\Pi = 0$, when the events (A_1, A_2) and (A_3,A_4) form $P_1 + P_2 - P_{12} = 1$ and $P_3 + P_4 - P_{34} = 1$ canceling each other. See Table 2. Figures 8 show probability diagrams. The shaded areas indicate non unique probabilities. The probability limits are obtained for $x \leq b$ from Figures 7, 8, and from (17a):

$$\Pi(x) = P(A_1) - P(A_1 \cap A_2) \qquad (18)$$

with A_1 is chosen first and $A_1 \cap A_2$ and A_2 follow, to obtain:

$$\Pi(x) \leq P(A_1) \leq 1$$
$$0 \leq P(A_1 \cap A_2) \leq 1 - \Pi(x)$$
$$0 \leq P(A_2) \leq 1 - \Pi(x) \qquad (19)$$

Similarly for A_3 chosen then $A_3 \cap A_4$ and A_4 follow on $b \leq x$, and from (17b):

$$\Pi(x) = 1 - [P(A_3) - P(A_3 \cap A_4)] \qquad (20)$$

producing:

$$1 - \Pi(x) \leq P(A_3) \leq 1$$
$$0 \leq P(A_3 \cap A_4) \leq \Pi(x)$$
$$0 \leq P(A_4) \leq \Pi(x) \qquad (21)$$

Table 2 has Method 2 summary, with $P_i=P(A_i)$, $P_{ij}=P(A_i \cap A_j)$, $i,j=1,2$. Note that the role of events A_1 and A_2 can be reversed, and similarly for A_3 and A_4, i.e. either pair of events can be chosen first. There may be some TBD probabilistic connection between events (A_1, A_2) and (A_3, A_4).

Table 2. $\Pi(x)$ using Method 2, Equations (17a,b)

x	A_1, A_3	A_2, A_4	A_1 vs. A_2, A_3 vs A_4

$0 \le x \le a$	$P_i = 0$	$P_j = 0$	$P_k + P_{k+1} = 0, k=1,3$
$a \le x \le b$ $P_3 = P_4 = 0$	A_1 non unique	A_2 non unique	$\Pi = P_1 - P_{12}$ $0 \le P_1 + P_2 - P_{12} \le 1$
$b \le x \le c$ $P_1 = P_2 = 0$	A_3 non unique	A_4 non unique	$\Pi = P_3 - P_{34}$ $0 \le P_3 + P_4 - P_{34} \le 1$
$c \le x$	$P_i = 1$	$P_j = 1$	$P_k + P_{k+1} - P_{k,k+1} = 1, k=1,3$

$0 \le x < a$	$P_i = 0$	$P_j = 0$	$P_k + P_{k+1} = 0, k=1,3$
$a \le x < b$ $P_3 = P_4 = 0$	A_1 non unique	A_2 non unique	$\Pi = P_1 + P_2 - P_{12}$ $0 \le P_1 + P_2 - P_{12} \le 1$
$b \le x < c$ $P_1 = P_2 = 0$	A_3 non unique	A_4 non unique	$\Pi = P_3 + P_4 - P_{34}$ $0 \le P_3 + P_4 - P_{34} \le 1$
$c \le x$	$P_i \le 1$	$P_j \le 1$	$P_k + P_{k+1} - P_{k,k+1} = 1, k=1,3$

3.4. Method 3

We use (1) and define fuzzy distribution decomposed as in (11), with the following choices for $F_1(x)$ and $F_2(x)$:

$$F_1(x) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \quad (22a)$$
$$F_2(x) = P(A_3 \cup A_4) = P(A_3) + P(A_4) - P(A_3 \cap A_4) \quad (22b)$$

Figures 7 applies here as well. For $b \le x$, the same process starts with two new events (A_3, A_4) with the corresponding $F_2(x)$ probabilities (22b). For $c \le x$ we have $\Pi(x) = 0$, when (A_1, A_2) and (A_3, A_4) form $P_1 + P_2 - P_{12} = 1$ and $P_3 + P_4 - P_{34} = 1$ compensating each other. See also Table 3. Figures 10 show two diagrams with the corresponding probabilities. The shaded areas indicate non unique probabilities. The probability limits similar to (14) and (16) can be obtained from Figures 7 and 8, as we assume A_1 is chosen first and then A_2 and $A_1 \cap A_2$ follow for $x \le b$:

$$0 \le P(A_1) \le \Pi(x)$$
$$0 \le \Pi(x) - P(A_1) \le P(A_2) \le \Pi(x)$$
$$0 \le P(A_1 \cap A_2) \le P(A_1) \le \Pi(x) \quad (23)$$

which is equivalent to:

$$0 \le P(A_1), P(A_2), P(A_1 \cap A_2) \le \Pi(x) \quad (24)$$

Similarly for A_3, A_4 and $A_3 \cap A_4$ for $b \le x$ we have:

$$0 \le P(A_3), P(A_4), P(A_3 \cap A_4) \le \Pi(x) \quad (25)$$

Table 3 has Method 3 summary. Note that the role of A_1 and A_2 can be reversed, and also for A_3 and A_4. We assume that the events A_1, A_2 are not related to A_3, A_4.
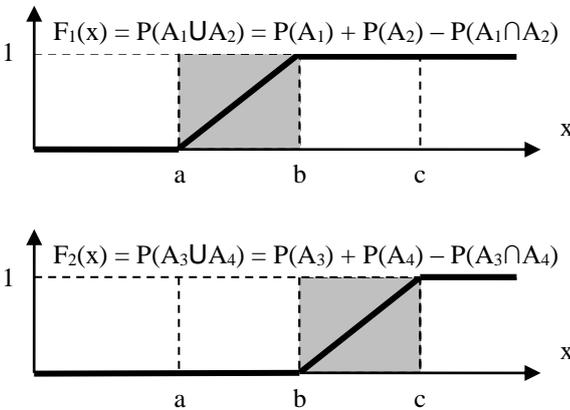
Figures 9. Method 3: CDF's $F_1(x)$ and $F_2(x)$

Table 3. $\Pi(x)$ using Method 3, Equations (22a,b)

x	A_1, A_3	A_2, A_4	A_1 vs. A_2, A_3 vs A_4

3.5. Other Methods

Before we state the main results in the next Section 4, few comments are in order. By considering Section 3.1 and Equation (11) we can attempt to decompose $\Pi(x)$ in other ways. For example an "obvious" choice is $\Pi(x) = F_1(x) - F_2(x) = P(A_1) - P(A_2)$. The problem with this choice is that it does not offer any probability variations (non uniqueness). Similarly if we choose $\Pi(x) = F_1(x) - F_2(x) = [P(A_1) + P(A_2)] - [P(A_3) + P(A_4)]$, the same comment applies, i.e. once we choose say $P(A_1)$, then $P(A_2)$ follows uniquely, and the same for $P(A_3)$ and $P(A_4)$. We comment on this further in Section 4.5.

4. Uncertainty Balance Principle

This section advances Section 4 results and states several general results. The main goal is to produce a usable and practical result to relate fuzzy and variable random data.

4.1. General Considerations

For simplicity we assume TFN within the intervals {a,b} and {b.c} non zero $\Pi(x)$. The results are general for any fuzzy distribution $\Pi(x)$ which can be represented by repeated procedure (11) for increasing x values. These distributions can be non convex, non normalized, and of other shapes, symmetric, non symmetric, unimodal and multimodal. Figure 10 shows a bimodal fuzzy distribution consisting of two non-overlapping TFNs. They can also overlap.
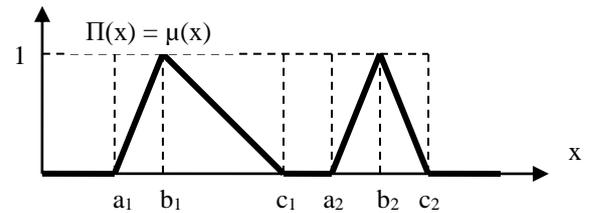
Figure 10. Bimodal TFN

The $\Pi(x)$ decomposition consists of two pairs of cumulative probabilistic distributions:

$$\Pi(x) = F_1(x) - F_2(x) + F_3(x) - F_4(x) \quad (26)$$

By an induction extension of (2), for "n" modal TFN we have the following general result:

Theorem 1. Fuzzy n-modal distribution function $\Pi(x)$ can be decomposed as a difference of sums of probabilistic cumulative distributions:

$$\Pi(x) = \sum F_i(x) - \sum F_j(x) \quad (27)$$

with i = 2k-1, j = 2k, k = 1,2,…,n, for any x, where the odd functions amount for rising portion of fuzzy distribution and even for the falling portion.

For an unimodal distribution, n = 1, (27) reduces to (11), and for a bimodal one, n = 2, and (27) reduces to (26). In terms of expressing (27) as specific probabilities, we have three possibilities depending on how we use (1).

4.2. Method 1

Using (26) and Section 3.2 results for bimodal case, we have:

$$\Pi(x) = P(A_1) - P(A_1 \cap A_2) + P(A_3) - P(A_3 \cap A_4) \quad (28a)$$

for two rising parts of $\Pi(x)$ and:

$$\Pi(x) = P(A_1 \cup A_2) - P(A_2) + P(A_3 \cup A_4) - P(A_4) \quad (28b)$$

for two falling parts of $\Pi(x)$. With (27) and (28a,b) we have the following:

Corollary 1.1. Fuzzy n-modal distribution function given in Theorem 1 can be further expressed as a difference of sums of probabilities:

$$\Pi(x) = \sum P(A_i) - \sum P(A_i \cap A_j) \quad (29a)$$

for two rising parts of $\Pi(x)$, and:

$$\Pi(x) = \sum P(A_i \cup A_j) - \sum P(A_j) \quad (29b)$$

for two falling parts of $\Pi(x)$, with i = 2k-1 and j = 2k, k = 1,2,…,n.

When n=1, for unimodal fuzzy distribution, (29a,b) reduce to (13) and (15), and for n=2, bimodal fuzzy distribution, (29a,b) reduce to (28a,b), respectively. Next, we define:

$$\sum \Delta P(A_i, A_j) = \sum \Delta P(A_i) + \sum \Delta P(A_j) \quad (30)$$

as total probability change, with:

$$\Delta P(A_k) = P(A_k)_M - P(A_k)_m \quad (31)$$

as probability range for event A_k, where "M" stands for maximum value, and "m" is for minimum value. We now state the following general result which relates a fuzzy distribution and a set of changes in related random event probabilities.

Theorem 2. Any multimodal fuzzy distribution $\Pi(x)$ can be expressed in terms of fuzzy presumption-invariant and x-invariant universal fuzzy-random Uncertainty Balance Principle:

$$\Pi(x) + \sum \Delta P(A_i, A_j) = 1 \quad (32a)$$
$$\Pi^*(x) \geq \sum \Delta P(A_i, A_j) \quad (32b)$$

for any x, with i = 2k-1, j = 2k, k = 1,2,…,n.

Note that practical implication of this Theorem is to be able to express fuzzy data distribution as a combination of a number of variable random events and corresponding probabilities. We illustrate this notion in Section 5 with numerical examples, and in particular in Example of Section 5.5 which discusses a specific fuzzy distribution and specific resulting variable probabilities. To continue, note that for simplicity, we did not burden the notation in Theorem 2 with stating dependency of ΔP's on x. The key feature of Theorem 2 is that it holds for any x and any presumption level of $\Pi(x)$. We prove the unimodal case when n=1, for TFN in Figure 2. The proof for any n and $\Pi(x)$ is straightforward, by repeating the procedure n times. From (14) and (16) we obtain (see also Section 5 examples):

For x ≤ b: (33)

$$P(A_1)_m = P(A_1)_M$$
$$\Delta P(A_1) = P(A_1)_M - P(A_1)_m = 0$$
$$P(A_2)_M = 1 - \Pi(x), \ P(A_2)_m = 0$$
$$\Delta P(A_2) = P(A_2)_M - P(A_2)_m = 1 - \Pi(x)$$

For b ≤ x: (34)

$$P(A_1)_m = \Pi(x), \ P(A_1)_M = 1$$
$$\Delta P(A_1) = P(A_1)_M - P(A_1)_m = 1 - \Pi(x)$$
$$P(A_2)_m = P(A_2)_X$$
$$\Delta P(A_2) = P(A_2)_M - P(A_2)_m = 0$$

Note that the point "b" is maximum $\Pi(x)$ point of a TFN, or any other unimodal fuzzy distribution. Replacing (33) and (34) into (32), for n = 1, we obtain:

$$\Pi(x) + \Delta P(A_1) + \Delta P(A_2) = 1 \quad (35)$$

or:

$$\Pi(x) + \Delta P(A_1, A_2) = 1 \quad (36a)$$
$$\Pi^*(x) \geq \Delta P(A_1, A_2) \quad (36b)$$

holding across the full range of argument x and $\Pi(x)$, and (36b) follows from (10).

One can interpret Theorem 2 result as "randomness" pool left to form fuzzy distribution to a random certainty. This also means that for higher "presumption" levels, near 1, corresponding randomness pool is smaller (less uncertainty to adjust) and for lower "presumption" levels it is larger (more uncertainty to adjust). Examples in Section 7 and Figures 10 show that clearly. We have the following result based on Theorem 2:

Corollary 2. Any fuzzy distribution derivative $d\Pi(x)/dx$ can be expressed for any argument x as a universal fuzzy-random Uncertainty Change Law:

$$d\Pi(x)/dx = - \sum d[\Delta P(A_i, A_j)]dx \quad (38a)$$
$$d\Pi^*(x)/dx \geq \sum d[\Delta P(A_i, A_j)]dx \quad (38b)$$

and i = 2k-1, j = 2k, k = 1,2,…,n. For n = 1, we have

$$d\Pi(x)/dx = - d[\Delta P(A_1, A_2)]/dx \quad (39a)$$
$$d\Pi^*(x)/dx \geq d[\Delta P(A_1, A_2)]/dx \quad (39b)$$

23

Due to the fact that the changes in two probabilities $\Delta P(A_1)$ and $\Delta P(A_2)$ are not zero at different arguments x (39a) reduces to a very simple fact:

$$d\Pi(x)/dx = - d[\Delta P(A_i)]/dx \qquad (40)$$

where $\Delta P(A_i)$ is $\Delta P(A_1)$ or $\Delta P(A_2)$ for i=2, depending on x value. Simply stated, (40) says that the change in fuzzy distribution is the opposite of probability change. This is also shown in Section 6 with numerical examples and in Figures 11. The first diagram shows $\Pi(x)$ changes with ΔP, for any fuzzy distribution. This is a consequence of Theorem 2 and presumption and x-invariant nature of it. The second diagram in Figure 11 indicates how $d\Pi/dx$ and $d(\Delta P)/dx$ relate, based on Corollary 2.1. The diagrams are universal for any $\Pi(x)$, for any $F_i(x)$. This is also illustrated in Section 6 with various distributions $\Pi(x)$.

4.3. Method 2

In this case Theorem 1 still holds as stated above. Instead of Corollary 1.1 and using (17ab), we have:
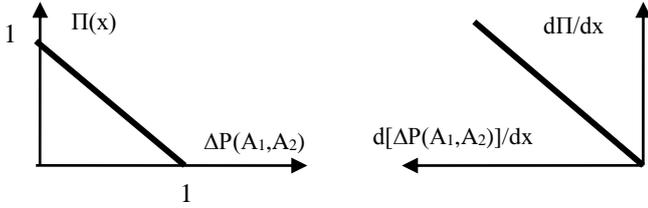
Figures 11: Method 1: Theorem 2 and Corollary 2

Corollary 1.2. Fuzzy n-modal distribution function given in Theorem 1 can be expressed as a difference of sums of probabilities:

$$\Pi(x) = \sum P(A_i) - \sum P(A_i \cap A_j) \qquad (41)$$

with i = 2k-1 and j = 2k, k = 1,2,…,n.

For n=1, we obtain $\Pi(x) = P(A_i) - P(A_i \cap A_j)$. Using (17a,b)-(21), with total probability change:

$$\sum \Delta P(A_i, A_j) = \sum \Delta P(A_i) - \sum \Delta P(A_i \cap A_j) \qquad (42)$$

we state the following general result similar to Theorem 2.

Theorem 3. Any multimodal fuzzy distribution $\Pi(x)$ can be expressed in terms of fuzzy presumption-invariant and x-invariant universal fuzzy-random Uncertainty Balance Principle. For any $x \le b$, we have:

$$\Pi(x) + \sum \Delta P(A_i, A_j) = 1 \qquad (43a)$$
$$\Pi^*(x) \ge \sum \Delta P(A_i, A_j) \qquad (43b)$$

and for any $b \le x$:

$$\Pi(x) = \sum \Delta P(A_i, A_j) \qquad (44a)$$
$$\Pi^*(x) + \sum \Delta P(A_i, A_j) \ge 1 \qquad (44b)$$

Similar to Theorem 2, the result above holds for any x and any presumption level of $\Pi(x)$. We prove the case when n=1, for TFN in Figure 2. The proof for any n and any $\Pi(x)$ is straightforward. From (18) through (21) we obtain:

For $x \le b$: $\qquad\qquad\qquad\qquad\qquad\qquad (45)$
$$P(A_1)_m = \Pi(x), \ P(A_1)_M = 1$$
$$\Delta P(A_1) = P(A_1)_M - P(A_1)_m = 1 - \Pi(x)$$
$$P(A_1 \cap A_2)_m = 0, \ P(A_1 \cap A_2)_M = 1 - \Pi(x)$$
$$\Delta P(A_1 \cap A_2) = P(A_1 \cap A_2)_M - P(A_1 \cap A_2)_m = 1 - \Pi(x)$$
$$P(A_2)_m = 0, \ P(A_2)_M = 1 - \Pi(x)$$
$$\Delta P(A_2) = P(A_2)_M - P(A_2)_m = 1 - \Pi(x)$$

Replacing (45) into (43), for n=1, we obtain:

$$\Pi(x) + \Delta P(A_1) + \Delta P(A_2) - \Delta P(A_1 \cap A_2) = 1 \qquad (46)$$

or:

$$\Pi(x) + \Delta P(A_1, A_2) = 1 \qquad (47a)$$
$$\Pi^*(x) \ge \Delta P(A_1, A_2) \qquad (47b)$$

For $b \le x$: $\qquad\qquad\qquad\qquad\qquad\qquad (48)$
$$P(A_3)_m = 1 - \Pi(x), \ P(A_3)_M = 1$$
$$\Delta P(A_3) = P(A_3)_M - P(A_3)_m = \Pi(x)$$
$$P(A_3 \cap A_4)_m = 0, \ P(A_3 \cap A_4)_M = \Pi(x)$$
$$\Delta P(A_3 \cap A_4) = P(A_3 \cap A_4)_M - P(A_3 \cap A_4)_m = \Pi(x)$$
$$P(A_4)_m = 0, \ P(A_4)_M = \Pi(x)$$
$$\Delta P(A_4) = P(A_4)_M - P(A_4)_m = \Pi(x)$$

Replacing (48) into (43), for n = 1, we obtain:

$$1 - \Pi(x) + \Delta P(A_3) + \Delta P(A_4) - \Delta P(A_3 \cap A_4) = 1 \qquad (49)$$

or:

$$\Pi(x) = \Delta P(A_3, A_4) \qquad (50a)$$
$$\Pi^*(x) + \Delta P(A_3, A_4) \ge 1 \qquad (50a)$$

Next, we have the following result based on Theorem 3:

Corollary 3. Any n-modal fuzzy distribution derivative $d\Pi(x)/dx$ can be expressed for any argument x as a universal fuzzy-random Uncertainty Change Law:

$$d\Pi(x)/dx = - \sum d[\Delta P(A_i, A_j)]/dx \qquad (51a)$$
$$d\Pi(x)/dx = \sum d[\Delta P(A_i, A_j)]/dx \qquad (51b)$$

for $x \le b$ and $b \le x$ respectively, i = 2k-1, j = 2k, k = 1,2,…,n.

For unimodal distribution, n = 1, Figures 12, we have:

$$d\Pi(x)/dx = - d[\Delta P(A_1, A_2)]/dx \qquad (52a)$$
$$d\Pi(x)/dx = d[\Delta P(A_3 A_4)]/dx \qquad (52b)$$

4.4. Method 3

Theorem 1 still holds. Using (22a) and (22b), we have:
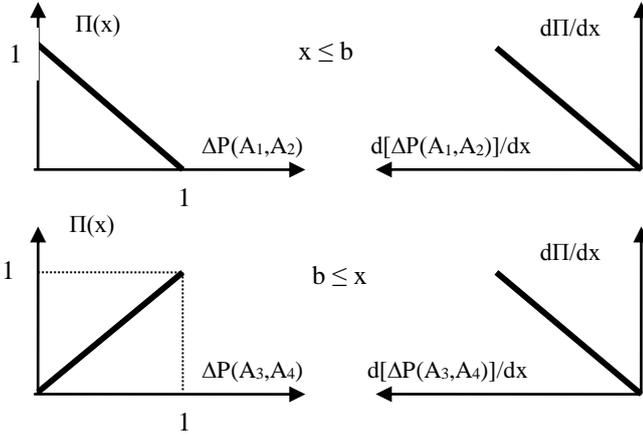
Figures 12: Method 2: Theorem 3 and Corollary 3

Corollary 1.3. Fuzzy n-modal distribution function given in Theorem 1 can be further expressed as a difference of sums of probabilities:

$$\Pi(x) = \sum P(A_i \cup A_j) = \sum P(A_i) + \sum P(A_j) - \sum P(A_i \cap A_j) \quad (54)$$

with i = 2k-1 and j = 2k, k = 1,2,…,n.

Using (23)-(25), as well as (18) and (21), and with:

$$\sum \Delta P(A_i, A_j) = \sum \Delta P(A_i) + \sum \Delta P(A_j) - \sum \Delta P(A_i \cap A_j) \quad (55)$$

for total probability change, we state the following:

Theorem 4. Any multimodal fuzzy distribution $\Pi(x)$ can be expressed in terms of fuzzy presumption-invariant and x-invariant universal fuzzy-random Uncertainty Balance Principle. For any $x \leq b$, we have:

$$\Pi(x) = \sum \Delta P(A_i, A_j) \quad (56a)$$
$$\Pi^*(x) + \sum \Delta P(A_i, A_j) \geq 1 \quad (56b)$$

and for any $b \leq x$:

$$\Pi(x) + \sum \Delta P(A_i, A_j) = 1 \quad (57a)$$
$$\Pi^*(x) \geq \sum \Delta P(A_i, A_j) \quad (57b)$$

We prove for n = 1, TFN in Figure 2. The proof for any n and any $\Pi(x)$ is straightforward. From (23)-(25) we obtain:

For $x \leq b$: (58)

$$P(A_1)_m = 0, \ P(A_1)_M = \Pi(x)$$
$$\Delta P(A_1) = P(A_1)_M - P(A_1)_m = \Pi(x)$$
$$P(A_1 \cap A_2)_m = 0, \ P(A_1 \cap A_2)_M = \Pi(x)$$
$$\Delta P(A_1 \cap A_2) = P(A_1 \cap A_2)_M - P(A_1 \cap A_2)_m = \Pi(x)$$
$$P(A_2)_m = 0, \ P(A_2)_M = \Pi(x)$$
$$\Delta P(A_2) = P(A_2)_M - P(A_2)_m = \Pi(x)$$

Replacing (58) into (56a), for n = 1, we obtain:

$$\Pi(x) = \Delta P(A_1) + \Delta P(A_2) - \Delta P(A_1 \cap A_2) \quad (59)$$

or:

$$\Pi(x) = \Delta P(A_1, A_2) \quad (60a)$$
$$\Pi^*(x) + \Delta P(A_1, A_2) \geq 1 \quad (60b)$$

with $\Delta P(A_1, A_2)$ representing total probability change for two events A_1 and A_2.

For $b \leq x$: (61)

$$P(A_3)_m = 0, \ P(A_3)_M = \Pi(x)$$
$$\Delta P(A_3) = P(A_3)_M - P(A_3)_m = \Pi(x)$$
$$P(A_3 \cap A_4)_m = 0, \ P(A_3 \cap A_4)_M = \Pi(x)$$
$$\Delta P(A_3 \cap A_4) = P(A_3 \cap A_4)_M - P(A_3 \cap A_4)_m = \Pi(x)$$
$$P(A_4)_m = 0, \ P(A_4)_M = \Pi(x)$$
$$\Delta P(A_4) = P(A_4)_M - P(A_4)_m = \Pi(x)$$

Replacing (61) into (56b), for n = 1, we obtain:

$$\Pi(x) + \Delta P(A_3) + \Delta P(A_4) - \Delta P(A_3 \cap A_4) = 1 \quad (62)$$

or:

$$\Pi(x) + \Delta P(A_3, A_4) = 1 \quad (63a)$$
$$\Pi^*(x) \geq \Delta P(A_3, A_4) \quad (63b)$$

with:

$$\Delta P(A_3, A_4) = \Delta P(A_3) + \Delta P(A_4) - \Delta P(A_3 \cap A_4) \quad (64)$$

representing total probability change of random events A_3 and A_4. We have the following result based on Theorem 4:

Corollary 4. Any n-modal fuzzy distribution derivative $d\Pi(x)/dx$ can be expressed for any argument x as a universal fuzzy-random Uncertainty Change Law:

$$d\Pi(x)/dx = \sum d[\Delta P(A_i, A_j)]/dx \quad (65a)$$
$$d\Pi(x)/dx = -\sum d[\Delta P(A_i, A_j)]/dx \quad (65b)$$

for $x \leq b$ and $b \leq x$ respectively, i=2k-1, j=2k, k=1,2,…,n. For unimodal distribution n=1, Figure 13, we have:

$$d\Pi(x)/dx = d[\Delta P(A_1, A_2)]/dx \quad (66a)$$
$$d\Pi(x)/dx = - d[\Delta P(A_3, A_4)]/dx \quad (66b)$$

and:

$$\Pi(x) = \Delta P(A_1, A_2) \quad (67b)$$
$$\Pi^*(x) + \Delta P(A_1, A_2) \leq 1 \quad (67a)$$

for $x \leq b$ and $b \leq x$ respectively.

Note that three methods (Theorems 2, 3, and 4) produce similar but not quite equivalent results. They offer different choices for variable probabilities. The common feature is that all state Uncertainty Balance Principle in which fuzzy distribution is either equal to probability change or offset by it, adding up to 1, i.e. certain event in random and presumption level 1 in fuzzy distribution. Corresponding Corollaries state Uncertainty Change Laws. Hence we have more than one option to transform fuzzy to random data, and vice versa. Section 7 illustrates all results with several numerical examples and suggests ideas how to generate fuzzy distributions using variable random events.
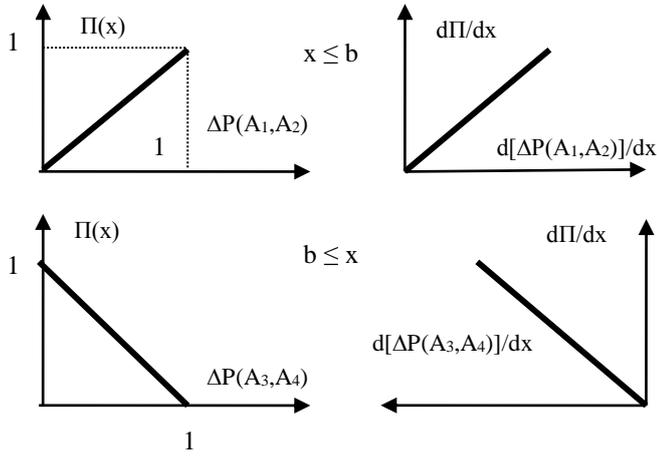
Figures 13. Method 3: Theorem 4 and Corollary 4

We also state the reverse result to Theorems 2, 3 and 4, i.e. random to fuzzy uncertainty alignment, as:

Corollary 5. Given a probabilistic and possibilistic uncertain variable X, with defined range of probabilities $\sum \Delta P(A_i,A_j)$ defined for random events A_i and A_j, i = 2k-1, j = 2k, k = 1,2,…,n, a possibilistic distribution $\Pi(x)$ can be formed observing Uncertainty Balance Principle in either of the forms:

$$\Pi(x) = 1 - \sum \Delta P(A_i,A_j) \qquad (68a)$$
$$\Pi(x) = \sum \Delta P(A_i,A_j) \qquad (68b)$$

depending on the uncertainty alignment method used, and the rising or falling side of $\Pi(x)$ represented.

Note that the results of this section can be also used in hard-soft data fusion where the staring point are probabilistic rather than possibilistic data. This gives our results an universal applicability in either fuzzy-random or random-fuzzy uncertainty data alignment.

4.5. Other Methods

As described in Section 4.5 we may consider to use other probability decomposition methods in (14) such as $\Pi(x) = F_1(x) - F_2(x) = [P(A_1) + P(A_2)] - [P(A_3) + P(A_4)]$. These choices do not offer any variability of probabilities, once one is chosen, the other follow uniquely. Hence there is no "fuzziness" involved, i.e. all $\Delta P(A_i,A_{i+1})$ are zero. For such cases Uncertainty Balance Principle reduces to "Certainty Principle" of the form:

$$\Pi(x) = P(A_1) + P(A_2) \qquad (69a)$$
$$\Pi(x) + P(A_3) + P(A_4) = 1 \qquad (69b)$$

for $x \leq b$ and $b \leq x$ respectively and it looks to be of less practical use. Note that for probability variability we need presence of non zero intersection of the random events and non zero probability $P(A_i \cap A_j)$. This is because specific $P(A_i \cap A_j)$ can be generated by a variety of random events A_i and A_j and their

corresponding probabilities. This is illustrated in Section 5.5.

4.6. Consistency Principle as Uncertainty Balance Principle

Referring back to Zadeh's Consistency Principle [2], as given in (7a) for an unimodal fuzzy distribution, one can re interpret it in the light of our Theorems 2, 3 and 4 which hold for any multimodal fuzzy distribution, any presumption level and any argument x. In this reinterpretation the Principle has a clear conceptual and numeric meaning, as well as an intuitive rationale. We can consider it as a "Fuzzy-Random Uncertainty Balance Principle" and it can be an alternative to Consistancy Principle given in (9). Instead of multiplying P_i and Π_i, we can use the sum of ΔP_i's and Π_i in the spirit of Theorems 2, 3 and 4 where "i" now points to a different x, that is x_i. Recall that all Theorems hold for any presumption level $\Pi(x)$ as well as any x. For example, from Theorem 2 we can redefine Consistence Principle using our Uncertainty Balance Principle general form $\Pi_i + \Delta P_i = 1$, equivalent to $\Pi_i = 1 - \Delta P_i = \Delta P^*_i$ or $\Pi^*_i \geq \Delta P_i$ which hold for any x_i, (i.e. on both sides of $\Pi(x)$) as:

$$\Gamma_X = \sum(\Pi_i + \Delta P_i) = \sum(\Delta P_i \Delta P^*_i) = n$$
$$\Gamma_X \leq \sum(\Pi_i \Pi^*_i) \qquad (70)$$

where i=1,…,n and ΔP_i is the corresponding total probability change in the x range where Π_i's are non zero. Complementary values are indicated with "*". Theorems 3 and 4 produce similar result, with slight modification:

$$\Gamma_X = \sum(\Pi_i \Pi^*_i) + \sum(\Delta P_i \Delta P^*_i) = n \qquad (71)$$

where i and j can go from 1 to n/2 or some other ratio. We see how Consistency Principle as defined above is a reflection of our Uncertainty Balance Principle, and besides an intuitiveness it has a definite numerical meaning as well. For example we can agree that $\Gamma_X = 10$ is better consistency than $\Gamma_X = 5$, if 10 and 5 are number of arguments x_i for which we have the agreement (or knowledge) that $\Pi_i = \Delta P_i$. This can be used in decision making situations when we need to combine soft (fuzzy) with hard (random) data, starting from either one. The assumption is, as stated earlier, that the uncertain variable X is both possibilistic as well as probabilistic. See also Table 4 bellow. We will elaborate on various applications of Uncertainty Balance Principle in our currently going research work on soft-hard or hard-soft data fusion.

4.7. Note On Soft-Hard Data Fusion

One of the practical motivations for this work, is to have a methodology to transform fuzzy data to random, and vice versa, so we can apply unique approach and available tools to both. In the first case it is probabilistic methodology, once fuzzy data are

26

described in terms of certain variable probabilities. At the same time we can reverse the process, and given probabilistic description of some phenomenon, where there is a natural variability of probabilities, we can transform random to fuzzy and use all the fuzzy tools available. In either case our approach can enhance decision making process where both soft and hard data are present, and both are to be used to make some decision. Section 3.6 points to one way to judge level of alignment of fuzzy and random data, in the situations where both are generated for a phenomenon which can be treated both as a fuzzy and as random. This phenomenon may come from a system of sensors or soft valuations such as coming from a human operator [21]. Table 4 has an intuitive summary of various equivalent descriptions and attributes found in literature on soft (human generated) and hard (sensor or machine) generated data. Other views on what is hard and soft and when to apply fuzzy vs. random are possible as well [41]-[44].

Table 4. Intuitive Soft and Hard Data Designations

Soft Data ↔ Human Operator	Hard Data ↔ Sensor (Sensors - Machine)
Subjective	Objective
Expresses Valuation	Expresses Measure
Qualitative	Quantitative
Possibilistic Methodology	Probabilistic Methodology
Fuzzy Models	Random Models
Distribution $\Pi_X(x) = \mu_X(x)$	Distribution $F_X(x)$

4.8. Other Fuzzy Distributions

Figures 14 show other types of fuzzy distributions which can be handled by our approach. The first one is a trapezoidal distribution which can be decomposed using a pair of CDFs. The second one is a bimodal and a combination of two distributions put together (gray area can belong to either). It can be decomposed by using two pairs of CDFs. The next one is a convex distribution with the maximum at "b". It can be decomposed by a pair of CDFs, with the break at "b". The last one is a concave distribution. First two fuzzy distributions consist of uniform random distributions, and the last two are not uniform. Any combination of the above distributions is possible too. Uncertainty Balance Principle and Uncertainty Change Law hold in any case, for uniform or non uniform distributions. In Section 6 we show four numerical examples, two uniform, two non uniform fuzzy distributions. Note that in every case the continuity conditions for CDFs are observed when $\Pi(x)$ is expressed in terms of Theorem 1.

5. Numerical Examples

5.1. Examples of $\Pi(x)$ Distributions

In this section we consider four numerical examples which illustrate the main results of the paper. Figure 2 and Equation (11) give a simple TFN decomposition with the CDFs as:

$$
\begin{aligned}
F_1(x) &= 0, & x < a \\
&= (x-a)/(b-a), & a \le x < b \\
&= 1, & b \le x \\
F_2(x) &= 0, & x < b \\
&= (x-b)/(c-b), & b \le x < c \\
&= 1, & c \le x
\end{aligned}
\tag{72}
$$
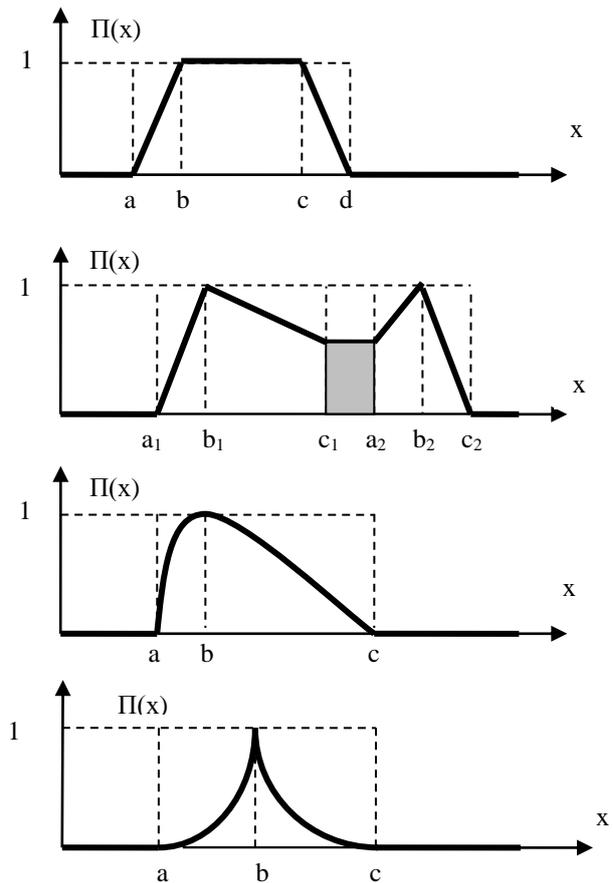
Figure 14. Various fuzzy distributions

which is used in Examples 1 and 2 bellow. Last two distributions in Figures 14 are used for Examples 3 and 4. Recall that F_1 and F_2 are equal to various probabilities as described in Methods 1, 2 and 3.

5.2. Method 1

Example 1 The symmetric TFN triplet {a, b, c} in Figure 2 is {2,3,4}. Table 5 has the values for x and the corresponding fuzzy "presumption" $\Pi(x)$ levels. The gray areas show $\Pi(x)$ and $\Delta P(A_2)$ and $\Delta P(A_1)$. For $x \le b$, the probability $P(A_1)$ is fixed for a fixed x. On the other hand, $P(A_2)$ resides in $\Delta P(A_2)$. We observe that for small $\Pi(x)$ values (low fuzzy "presumption") the corresponding range of $P(A_2)$ is wider (more uncertainty), and for bigger values of $\Pi(x)$ (high fuzzy "presumption" level), range of $P(A_2)$ is narrower (less uncertainty). This makes intuitive sense. We have the

27

same situation for b ≤ x, except that the non unique probability is now $P(A_1)$ residing in $\Delta P(A_1)$.

Example 2. We change the triplet {a,b,c} in Figure 5 to {10,15,30}, a non symmetric TFN with a larger spread of x. Table 6 shows numerical values. The same comments apply as in Example 1. Note that $\Pi(x)$ values and probability ranges $\Delta P(A)$ are as in Example 1 (confirming x-and fuzzy presumption invariance).

Example 3. For Example 3 we choose a fuzzy distribution described by a half circle with {a,b,c} = {1,2,3} where "b" is at the circle center, with radius 1, and outside of {1,2,3} distribution is 0. For $1 \le x < 3$:

$$\Pi(x) = \sqrt{[1 - (x - 2)^2]} \qquad (73)$$

Table 6 has the results. Note that the $\Pi(x)$ values are not uniformly distributed. The distribution changes the fastest right from x=1 and left from x=3, as in Table 7. Still linear Uncertainty Balance Principle holds.

Example 4. This example is two quarter circles of radius 1, centered at (1,1) and (3,1):

$$\Pi(x) = 1 - \sqrt{[1 - (x - 1)^2]}, \quad 1 \le x < 2$$
$$= 1 - \sqrt{[1 - (x - 3)^2]}, \quad 2 \le x < 3 \qquad (74)$$

$\Pi(x)$ values are not uniformly distributed. The distribution changes faster near x=2, on both sides. For all examples Figures 12, 13 and 14 as illustrated in Tables 5, 6, 7 and 8, confirm linear relationships of any $\Pi(x)$ (or $\Pi^*(x)$) with total probability change $\Delta P(A_i,A_j)$, as well as linear relationship of $d\Pi(x)/dx$ and $d[\Delta P(A_1,A_2)]/dx$.

Table 5. Example 1

x	F₁	F₂	Π	P₁M	P₁m	ΔP₁	P₂m	P₂M	ΔP₂
1.5	0	0	0	0	0	0	0	1	1
2	0	0	0	0	0	0	0	1	1
2.2	0.2	0	0.2	0.2	0.2	0	0	0.8	0.8
2.4	0.4	0	0.4	0.4	0.4	0	0	0.6	0.6
2.6	0.6	0	0.6	0.6	0.6	0	0	0.4	0.4
2.8	0.8	0	0.8	0.8	0.8	0	0	0.2	0.2
3	1	0	1	1	1	0	0	0	0
3.1	1	0.1	0.9	1	0.9	0.1	0.1	0.1	0
3.3	1	0.3	0.7	1	0.7	0.3	0.3	0.3	0
3.5	1	0.5	0.5	1	0.5	0.5	0.5	0.5	0
3.7	1	0.7	0.3	1	0.3	0.7	0.7	0.7	0
4	1	1	0	1	0	1	1	1	0
4.5	1	1	0	1	0	1	1	1	0

Table 6. Example 2

x	F₁	F₂	Π	P₁M	P₁m	ΔP₁	P₂m	P₂M	ΔP₂
8	0	0	0	0	0	0	0	1	1
10	0	0	0	0	0	0	0	1	1
11	0.2	0	0.2	0.2	0.2	0	0	0.8	0.8
12	0.4	0	0.4	0.4	0.4	0	0	0.6	0.6
13	0.6	0	0.6	0.6	0.6	0	0	0.4	0.4
14	0.8	0	0.8	0.8	0.8	0	0	0.2	0.2
15	1	0	1	1	1	0	0	0	0
18	1	0.2	0.8	1	0.8	0.2	0.2	0.2	0
21	1	0.4	0.6	1	0.6	0.4	0.4	0.4	0
27	1	0.8	0.2	1	0.2	0.8	0.8	0.8	0
30	1	1	0	1	0	1	1	1	0
35	1	1	0	1	0	1	1	1	0

Table 7. Example 3

x	F₁	F₂	Π	P₁M	P₁m	ΔP₁	P₂m	P₂M	ΔP₂
.5	0	0	0	0	0	0	0	1	1
1	0	0	0	0	0	0	0	1	1
1.2	0.6	0	0.6	0.6	0.6	0	0	0.4	0.4
1.4	0.8	0	0.8	0.8	0.8	0	0	0.2	0.2
1.6	0.9	0	0.9	0.9	0.9	0	0	0.1	0.1
1.8	.98	0	.98	.98	.98	0	0	0.02	0.02
2	1	0	1	1	1	0	0	0	0
2.2	1	0.6	.98	1	0.98	0.02	0.6	0.6	0
2.4	1	0.8	0.9	1	0.9	0.1	0.8	0.8	0
2.8	1	.98	0.6	1	0.4	0.4	.98	.98	0
3	1	1	0	1	0	1	1	1	0
3.5	1	1	0	1	0	1	1	1	0

Table 8. Example 4

x	F₁	F₂	Π	P₁M	P₁m	ΔP₁	P₂m	P₂M	ΔP₂
.5	0	0	0	0	0	0	0	1	1
1	0	0	0	0	0	0	0	1	1
1.2	.02	0	.02	.02	.02	0	0	.98	.98
1.4	.08	0	.08	.08	.08	0	0	.92	.92
1.6	.2	0	.2	.2	.2	0	0	.8	.8
1.8	.4	0	.4	.4	.4	0	0	.6	.6
2	1	0	1	1	1	0	0	0	0
2.2	1	.02	.98	1	.98	.02	.02	.02	0
2.4	1	.08	.92	1	.92	.08	.08	.08	0
2.6	1	.2	.8	1	.8	.2	.2	.2	0
2.8	1	.4	.6	1	.6	.4	.4	.4	0
3	1	1	0	1	0	1	1	1	0
3.5	1	1	0	1	0	1	1	1	0

5.3. Method 2

We use Example 1 from Method 1 again. Now we have more complexity due to more variable probabilities, $P(A_1)$ and $P(A_2)$ but also $P(A_1 \cap A_2)$, which is not always zero. This gives more options to form fuzzy distribution $\Pi(x)$. Conditions (19), (20) and (21) are used to determine values of various probabilities, as summarized in Tables 9.1 and 9.2 bellow. For simplicity we only included minimum number of values of fuzzy distribution $\Pi(x)$, due to many different combinations of individual probabilities $P(A_1)$, $P(A_2)$, and $P(A_1 \cap A_2)$. Table 9.1 corresponds to $x \le b$, rising part of $\Pi(x)$ and $F_1(x)$, and Table 9.2 corresponds to $b \le x$ and the falling part of $\Pi(x)$ and $F_2(x)$. Table 9.1 confirms Theorem 3 and Corollary 3. Of many options in Tables 9 we can simplify things by choosing, for example, various probabilities to be 1 or 0 at the critical points [a,b,c] (**boldfaced**). Also, between the critical points we can reduce number of options. Double lines indicate breaks in Tables 9 where probability values in between are obvious. Examples 2, 3 and 4, are not repeated for simplicity.

5.4. Method 3

For simplicity we will not repeat all the details for Method 3. The key is for the probabilities to follow conditions (23) and (25), given in Theorem 4 proof. Other comments given for Method 3 apply for Method 2 as well. As in previous two Methods, we can see a variety of probability choices which generate the same presumption level $\Pi(x)$. Depending on the specific application we can choose specific ΔP_i's.

28

Table 9.1 Example 1 (F₁, P₁ₘ, P₂ₘ and P₁₂ₘ are zero)

x	Π	Π*	F₁	P₁M	ΔP₁	P₂M	ΔP₂	P₁₂M	ΔP₁₂
0.5	0	1	0	0	0	0	0	0	0
1	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
1.2	0.2	0.8	0.2	0	0	0.2	0.2	0	0
				0.1	0.1	0.1	0.1	0	0
				0.2	0.2	0	0	0	0
				0.2	0.2	0.1	0.1	0.1	0.1
				0.2	0.2	0.2	0.2	0.2	0.2
1.5	0.5	0.5	0.5	0.0	0.0	0.5	0.5	0	0
				0.1	0.1	0.4	0.4	0	0
				0.2	0.2	0.3	0.3	0	0
				0.4	0.4	0.1	0.1	0	0
				0.5	0.5	0.0	0.0	0	0
				0.5	0.5	0.1	0.1	0.1	0.1
				0.5	0.5	0.5	0.5	0.5	0.5
2	**1**	**0**	**1**	**0.0**	**0.0**	**1.0**	**1.0**	**0**	**0**
				0.1	0.1	0.9	0.9	0	0
				0.2	0.2	0.8	0.8	0	0
				0.3	0.3	0.7	0.7	0	0
				0.8	0.8	0.2	0.2	0	0
				0.9	0.9	0.1	0.1	0	0
				1.0	**1.0**	**0.0**	**0.0**	**0**	**0**
				1.0	1.0	0.1	0.1	0.1	0.1
				1.0	1.0	0.9	0.9	0.9	0.9
				1.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**

Table 9.2 Example 1 (F₂, P₃ₘ, P₄ₘ and P₃₄ₘ are zero)

x	Π	Π*	F₂	P₃M	ΔP₃	P₄M	ΔP₄	P₃₄M	ΔP₃₄
2.1	0.9	0.1	0.1	0.1	0.1	0	0	0	0
				0.1	0.1	0.1	0.1	0.1	0.1
2.2	0.8	0.2	0.2	0	0	0.2	0.2	0	0
				0.1	0.1	0.1	0.1	0	0
				0.2	0.2	0.1	0.1	0.1	0.1
				0.2	0.2	0.2	0.2	0.2	0.2
2.5	0.5	0.5	0.5	0	0	0.5	0.5	0	0
				0.1	0.1	0.4	0.4	0	0
				0.1	0.1	0.5	0.5	0.1	0.1
				0.2	0.2	0.3	0.3	0	0
				0.2	0.2	0.5	0.5	0.2	0.2
				0.3	0.3	0.2	0.2	0	0
				0.3	0.3	0.3	0.3	0.1	0.1
				0.3	0.3	0.5	0.5	0.3	0.3
				0.4	0.4	0.1	0.1	0	0
				0.4	0.4	0.1	0.1	0.1	0.1
				0.4	0.4	0.4	0.4	0.4	0.4
				0.5	0.5	0	0	0	0
				0.5	0.5	0.0	0.0	0.1	0.1
				0.5	0.5	0.5	0.5	0.5	0.5
3.0	**0**	**1**	**1**	**0**	**0**	**1.0**	**1.0**	**0**	**0**
				0.1	0.1	0.9	0.9	0	0
				0.1	0.1	1.0	1.0	0.1	0.1
				0.2	0.2	0.8	0.8	0	0
				0.2	0.2	0.9	0.9	0.1	0.1
				0.2	0.2	1.0	1.0	0.2	0.2
				0.3	0.3	0.7	0.7	0	0
				0.4	0.4	0.9	0.9	0.3	0.3
				0.4	0.4	1.0	1.0	0.4	0.4
				0.5	0.5	0.5	0.5	0	0
				0.8	0.8	0.2	0.2	0	0
				0.9	0.9	0.1	0.1	0	0
				1.0	**1.0**	**0**	**0**	**0**	**0**
				1.0	1.0	0.1	0.1	0.1	0.1
				1.0	1.0	0.9	0.9	0.9	0.9
				1.0	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
3.5	0	1	1	1.0	1.0	1.0	1.0	1.0	1.0

5.5. Practical Example of Π(x)

Finally, we illustrate our methodology by a specific TFN fuzzy distribution describing people productivity, and other similar fuzzy applications, [23]-[29]. The triplet {a,b,c} is {10,45,90} expressed in years. Assumption is that an average person is the most productive around age of 45 which corresponds to Π(x) = 1. On the opposite end, it is assumed person has zero productivity at age of 10 and 95, hence Π(x) =0. Obviously this is just an approximation but it serves our purposes here. Using (71) we obtain Table 10 where the second row will be described shortly.

Table 10. People Productivity Example, Method 2

x, Age	Π	Π*	ΔP₁	ΔP₂	ΔP₁₂	ΔP₃	ΔP₄	ΔP₃₄
			Educ	Empl	Edu. x Empl	Health	Marri	Health x Marri
10	0	1	1	1	1	0	0	0
27.5	0.5	0.5	0.5	0.5	0.5	0	0	0
36.25	0.75	0.25	0.25	0.25	0.25	0	0	0
45	1	0	0	0	0	1	1	1
57.5	0.75	0.25	0.25	0.25	0.25	0.75	0.75	0.75
70	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
95	0	1	0	0	0	0	0.5	0

If we use Method 2, the following relations hold:

For x ≤ 45: $\Pi(x) + \Delta P(A_1) + \Delta P(A_2) - \Delta P(A_1 \cap A_2) = 1$
For 45 ≤ x: $\Pi^*(x) + \Delta P(A_1) + \Delta P(A_2) - \Delta P(A_1 \cap A_2) = 1$

So how do we interpret the results ? It is assumed the variable X is both possibilistic (fuzzy) and probabilistic (random). Data produced by different sources related to peoples productivity may be soft (subjective, expert opinions) or hard (statistics, random analysis, objective). Hence we can assume that the fuzzy presumption level Π(x) is generated by an interplay of $P(A_i)$ and $P(A_i \cap A_{i+1})$ per of Corollary 1.2, i.e.:

For x ≤ 45: $\Pi(x) = P(A_1) - P(A_1 \cap A_2)$
For 45 ≤ x: $\Pi(x) = P(A_3) - P(A_3 \cap A_4)$

From Theorem 3 we know that all of the probabilities vary for a given presumption level Π(x) hence producing "fuzziness" of uncertain variable X. Table 10 indicates that as well. Next we can assume that the random (probabilistic) events A_1 through A_2 are related to level of people's productivity. For example, given the age x and presumption level Π(x) we can assume there are other factors playing the role in productivity for the given age. One possibility may be as given in Table 11. There are other posibilities as well, per specific interest. Second row in Table 10 indicates choices from Table 11.

Table 11. Random event categories

	A₁ ΔP₁	A₂ ΔP₂	A₁₂ ΔP₁₂	A₃ ΔP₃	A₄ ΔP₄	A₃₄ ΔP₃₄
Education	X					
Employment		X				
Educ. vs. Employm.			X			
Health				X		
Marriage					X	
Health vs. Marriage						X

All of the A_i's and their probabilities ΔP's can be described by some CDF, uniform or non uniform,

29

obtained analytically or statistically. Once A_i are defined we can interpret variable probabilities accordingly. For example for x=36.25 years in Table 11, presumption level is $\Pi(x) = 0.75$. Corresponding ΔP's are all 0.25 on rising part of $\Pi(x)$ and that can be interpreted as variability of education, employment and their intersection for that age. On the other hand for the falling part of $\Pi(x)$, when $\Pi(x)=0.75$ we note larger variability of 0.75 of ΔP's, and this refers to health, marriage and their intersection. Methods 1 and 3 could have been used as well. Which Method do we use may depend on a specific fuzzy distribution and application, and more research needs to be done in this subject, looking into specific applications. Note also that Big Data methodology or various statistical methods can be used to have specific choices for the probabilities in Table 11, which exhibit strongest correlations with the original uncertain soft data.

Per Corollary 5, one can reverse the problem. Namely, instead of starting with the possibilistic description and produce (or interpret it via) a probabilistic one, we can start with the probabilistic and produce possibilistic description using essentially the same methodology described in this paper, just in a reverse order. This may be advantageous in some specific cases where we have the fuzzy tools or other fuzzy applications available, but the initial data is random and probabilistic in nature.

6. Uncertainty Alignment Algorithms

In this Section we present pseudo codes for the algorithms for generating various probability values in Methods 1, 2 and 3, once fuzzy distribution $\Pi(x)$ is specified. We assume in each case that the uncertain variable X is both possibilistic and probabilistic, hence Uncertainty Balance and Consistency Principles apply. We do not elaborate on how to generate various probabilities, which is specific for an application at hand, such as illustrated in Section 5.5

Method 1 (Section 3.2)
START Method 1
   For Rising Part: $(x \leq b)$
        Given $\Pi(x)$, $0 \leq \Pi(x) \leq 1$
        $\Pi(x) = F_1(x) - F_2(x) = P(A_1) - P(A_1 \cap A_2)$
          $P(A_1 \cap A_2) = 0$
        Choose Random Event A_1
          Set $P(A_1) = \Pi(x)$
        Choose Random Event A_2
          $P(A_2)$, $P(A_1 \cap A_2) \leq P(A_2) \leq 1 - \Pi(x)$
   End Rising Part
   For Falling Part: $(b \leq x)$
        Given $\Pi(x)$, $0 \leq \Pi(x) \leq 1$
        $\Pi(x) = F_1(x) - F_2(x) = P(A_1 \cup A_2) - P(A_2)$
          $P(A_1 \cup A_2) = 1$
        Choose Random Event A_2
          $P(A_2) = 1 - \Pi(x)$
        Choose Random Event A_1
          $P(A_1) \rightarrow \Pi(x) \leq P(A_1) \leq P(A_1 \cup A_2)$
   End Falling Part
END Method 1

Method 2 (Section 3.3)
START Method 2
   For Rising Part: $(x \leq b)$
        Given $\Pi(x)$, $0 \leq \Pi(x) \leq 1$
        $\Pi(x) = F_1(x) = P(A_1) - P(A_1 \cap A_2)$
          $P(A_1 \cap A_2) \neq 0$
        Choose Random Event A_1
          $P(A_1) \rightarrow \Pi(x) \leq P(A_1) \leq 1$
        Choose Random Event A_2
          $P(A_2) \rightarrow 0 \leq P(A_2) \leq 1 - \Pi(x)$
          $P(A_1 \cap A_2) \rightarrow 0 \leq P(A_1 \cap A_2) \leq 1 - \Pi(x)$
   End Rising Part
   For Falling Part: $(b \leq x)$
        Given $\Pi(x)$, $0 \leq \Pi(x) \leq 1$
        $\Pi(x) = 1 - F_2(x) = P(A_3) - P(A_3 \cap A_4)$
          $P(A_3 \cap A_4) \neq 0$
        Choose Random Event A_3
          $P(A_3) \rightarrow \Pi(x) \leq P(A_3) \leq 1$
        Choose Random Event A_4
          $P(A_4) \rightarrow 0 \leq P(A_4) \leq 1 - \Pi(x)$
          $P(A_3 \cap A_4) \rightarrow 0 \leq P(A_3 \cap A_4) \leq 1 - \Pi(x)$
   End Falling Part
END Method 2

Method 3 (Section 3.4)
START Method 3
   For Rising Part: $(x \leq b)$
        Given $\Pi(x)$, $0 \leq \Pi(x) \leq 1$
        $\Pi(x) = F_1(x) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$
          $P(A_1 \cap A_2) \neq 0$
        Choose Random Event A_1
          $P(A_1) \rightarrow 0 \leq P(A_1) \leq \Pi(x)$
        Choose Random Event A_2
          $P(A_2) \rightarrow 0 \leq P(A_2) \leq \Pi(x)$
          $P(A_1 \cap A_2) \rightarrow 0 \leq P(A_1 \cap A_2) \leq \Pi(x)$
   End Rising Part
   For Falling Part: $(b \leq x)$
        Given $\Pi(x)$, $0 \leq \Pi(x) \leq 1$
        $\Pi(x) = 1 - F_1(x) = P(A_3) + P(A_4) - P(A_3 \cap A_4)$
          $P(A_3 \cap A_4) \neq 0$
        Choose Random Event A_3
          $P(A_3) \rightarrow 0 \leq P(A_3) \leq \Pi(x)$
        Choose Random Event A_4
          $P(A_4) \rightarrow 0 \leq P(A_4) \leq \Pi(x)$
          $P(A_3 \cap A_4) \rightarrow 0 \leq P(A_3 \cap A_4) \leq \Pi(x)$
   End Falling Part
END Method 3

7. Conclusion
In this paper we define new fuzzy to random uncertainty alignment methodology, in which fuzziness can be described as precisely defined non unique randomness. We employ the most basic properties of random and fuzzy distributions for this result, starting from fuzzy distributions decomposed as a combination of probabilistic cumulative distribution functions, CDFs, rather than probabilistic density functions, PDF's, which may not always exist. We also give precise both upper and lower bounds of changes in random distributions, required to produce data fuzziness. The range of randomness of the corresponding probabilistic events is a function of fuzzy distribution presumption levels and it holds for

any fuzzy distribution. The main results is a universal fuzzy-random (possibilistic-probabilistic) uncertainty alignment law which we named Uncertainty Balance Principle, for its simple statement $\Pi(x) + \sum \Delta P(A_i, A_j) = 1$, which is a linear law, fuzzy presumption-invariant and fuzzy argument x invariant for any fuzzy distribution. Another byproduct of this Principle is also a linear law, Uncertainty Change Law which relates changes in fuzzy distribution against corresponding changes in probabilities. Our results hold for any fuzzy distributions, triangular, trapezoidal, convex, non convex, symmetric or not, normalized or not, uni modal or multi modal alike. This universal range of applicability comes as a result of employing CDF rather than PDF. The results of this paper can be employed effectively in a variety of data fusion and decision problems where both objective (hard, random, probabilistic, sensor based) data are to be fused with subjective (soft, fuzzy, possibilistic, human based) data [9], [10], [21]. They can be also used to generate random from fuzzy data for other applications. Additional feature of our approach is a reverse applicability of the results, i.e. going from random to fuzzy. If the range of probabilities is given we can form

corresponding fuzzy distribution satisfying Uncertainty Balance principle. Another way to interpret the results is as precise mathematical description of Consistency Principle first introduced by Zadeh in his classic "possibility" paper [2], as a loose and intuitive notion connecting fuzzy and random data. As defined in [2] this Principle relates fuzzy and random distributions in an intuitive way. In our paper, a precise mathematical definition is given for the modified Consistency Principle in terms of Uncertainty Balance Principle of this paper, as a measure of agreement between fuzzy and random data distributions. This can be effectively used to measure level of agreement between related fuzzy and random data in decision making process. In our future work we will extend the results in decision making areas such as machine-human data fusion where using both types of data is crucial for the fusion usefulness. Also, further properties of both random and fuzzy data will be analyzed in the light of the paper's main results. In particular we will consider relationship between probabilistic and possibilistic axioms in the light of this paper results, as well as special cases such as random events independence or dependence.

8. References

[1] L. A. Zadeh, Is there a need for fuzzy logic? Elsevier, Info. Sciences, 178, 2751–2779, 2008.

[2] L. A. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1, 3-28, 1978.

[3] L. A. Zadeh, Possibility theory and soft data analysis, Selected papers by Lotfi Zadeh, 515-541, 1981.

[4] L. A. Zadeh, Fuzzy sets, Information and Control, 8, 338-353, 1965.

[5] L. A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Information Sciences, Part I: 8, 199-249, Part II: 8, 301-357; Part III: 9, 43-80, 1975.

[6] R. Bellman and L. Zadeh, Decision-Making in a fuzzy environment, Management Science, 17, No. 4, 1979.

[7] A. Kaufmann and M. M. Gupta, Introduction to Fuzzy Arithmetic, Theory and Applications, Van Nost. Reinhold, 1985.

[8,] H. J. Zimmermann, Fuzzy Set Theory and its Applications, 4th Edition, Kluwer Academic Publ., 2001.

[9] D. Dubois, Possibility Theory and Statistical Reasoning, Institut de Recherche en Informatique de Toulouse, May, 2006

[10] D. Dubois and H. Prade, Possibility theory, Probability theory and multiple valued logics: A clarification, Annals of Mathematics and Artificial Intelligence, 32, 35-66, 2002.

[11] D. Dubois, H. Prade and P. Smets, New Semantics for Quantitative Possibility Theory, 2nd Int'l Symp. on Imprecise Probabilities and Their Applications, Ithaca, N.York, 2001.

[12] D. Dubois and H. Prade, Unfair coins and necessity measures: towards a possibilistic interpretation of histograms, Fuzzy Sets and Systems, 10, 15-20, 1983.

[13] D. Dubois and H. Prade, Fuzzy sets and statistical data, European J. of Operations Research, 25, 345-356, 1986.

[14] D. Dubois and H. Prade, The mean value of a fuzzy number, Fuzzy Sets and Systems, 24, 279-300, 1987.

[15] D. Dubois and H. Prade, Possibility Theory, Plenum Press, New York, 1988.

[16] D.Dubois and H.Prade, When upper probabilities are possibility measures, Fuzzy Sets and Systems, 49, 65-74, 1992.

[17] R. van der Helm, Towards a clarification of probability, possibility and plausibility: How semantics could help futures practice to improve, Foresight, 8, No.3, 17-27, 2008.

[18] P. Agarwal and H.S. Najal, Possibility theory vs possibility theory in fuzzy measure theory, P. Agarwal Int'l J.of Engineering Research and Applications, 5, No.5, 37-43, 2015.

[19] F.P.A. Coolen et all, Imprecise probability, Int'l Encyclopedia of Statistical Science, Springer, 2010.

[20] W. Eschenbach, Triangular Fuzzy Numbers and the IPCC, February, 2012.

[21] M. P. Jenkins et all, Towards context aware data fusion: Modeling and integration of situationally qualified human observations to manage uncertainty in a hard-soft fusion process, Information Fusion, 21, 130-144, 2015.

[22] Creating Membership Functions (PID, Fuzzy Logic Toolkit) LabVIEW, National Instruments, 2011.

[23] B. Kaur, M. Bala, M. Kumar, Comparitive Analysis of Fuzzy based Wildfire Detection

Techniques, Int'l J. of Scientific & Engineering Research, Volume 5, Issue 7, July, 2014.

[24] S. Şentürk, Fuzzy Regression Control Chart Based on α-cut Approximation, International Journal of Computational Intelligence Systems, 3, No. 1, 123-140, April 2010.

[25] O. B. Onuwa, Fuzzy expert system for malaria diagnosis, Oriental J. of Computer Science and Technology, 7, No. 2, 273-284, June 2014.

[26] P. Babashamsi, A. Golzadfar, N. I. Yusoff, H. Ceylan, N. G. Nor, Integrated Fuzzy Analytic Hierarchy Process and VIKOR Method in the Prioritization of Pavement Maintenance Activities, Int'l J. of Pavement Research and Technology, 2016.

[27] J. M. Garibaldi, R. I. John, Choosing Membership Functions of Linguistic Terms, 2016.

[28] E. Raufaste and Rui Da Silva Neves, Empirical Evaluation of Possibility Theory in Human Radiological Diagnosis, 13th Euro. Conference on Artificial Intelligence, Edited by H. Prade, John Wiley and Sons, 1998.

[29] I. Iancu, A Mamdani Type Fuzzy Logic Controller, Fuzzy Logic - Controls, Concepts, Theories and Applications, Edited by E. Dadios, In Tech, March, 2012.

[30] M. S. Yang, M. C. Liu, On possibility analysis of fuzzy data, Elsevier, Fuzzy sets and systems, 94, 171-183, 1998.

[31] Y. Narukawa, V. Torra, T. Gakuen, Fuzzy measure and probability distributions: Distorted probabilities, 2016.

[32] G. de Cooman, Possibility theory 1, The measure- and integral-theoretic groundwork, Universiteit Gent, Vakgroep Elektrische Energietechniek, 1996.

[33] B. Liu, Why is There a Need for Uncertainty Theory ?, J. of Uncertain Systems, 6, 1, 3-10, 2012.

[34] J. L. Doob, Stochastic Processes, J. Willey and Sons, 1953.

[35] W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 1 and 2, J. Willey and Sons, 1950.

[36] Gilles Mauris, Possibility distributions: A unified representation of usual direct-probability-based parameter estimation methods, Int'l Journal of Approximate Reasoning, 52, 1232–1242, 2011.

[37] L. Sanchez, J. Casillas, O. Cord, M. Jose del Jesus, Some relationships between fuzzy and random set-based classifiers and models, International Journal of Approximate Reasoning, 29, 175–213, 2002.

[38] Arnold F. Shapiro, Fuzzy random variables, Insurance: Mathematics and Economics, 44, 307-314, 2009.

[39] Sevil Sentuik, Fuzzy regression control chart based on α-cut approximation, Int'l Journal on Computational Intelligence Systems, 3, No.1, 123-140, April 2010.

[40] M. Hodzic, Fuzzy to Random Uncertainty Alignment, Southeast Europe Journal of Soft Computing, 5, No. 1, 58-66, March 2016.

[41] I. Couso et all, Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables: An Introd. for Ph.D. Students and Practitioners, Springer, 2014.

[42] B. de Finetti Theory of probability (2 vols.), J. Wiley & Sons, Inc., New York, 1974.

[43] J.M. Bernardo, Reference analysis, Handbook of statistics, 25, 17-90, 2005.

[44] E.T. Jaynes,"Bayes Methods:General Background" Maximum-Entropy & Bayesian Methods in Applied Statistics, J. H. Justice (ed.). Cambridge U. Press, 1986.

A review on Nanoparticle and Protein interaction in biomedical applications

Jasmin Šutković

International University of Sarajevo, Hrasnickacesta
15,Ilidza,Sarajevo,Bosnia and Herzegovina
jsutkovic@ius.edu.ba

Amina Jašarević

International University of Sarajevo, Hrasnickacesta 15,Ilidza,
Sarajevo, Bosnia and Herzegovina
amina_jasharevicka@hotmail.com

Abstract

Nanoparticles are molecules with size depended chemical and pyhsical characteristics, enabling interesting and correlated approaches while dealing with fundamental biological questions. Nanoparticles are capable of strong and important interaction with other molecules. Many different nanoparticles are produced, with variety of different roles, but Gold nanoparticle as metal based beads, have specific importance due to their attractive physical and chemical properties, biocompatibility, and facile surface modification. In general, nanoparticles have the ability to interact with whole physiological surrounding once when they enter human body. In most of the cases, first molecule they interact with are proteins, which are the main constituens of human body and the driving force of most of the biological processes. This understanding of interaction between nanoparticles and proteins represents an important essence for secure and efficient application of nanoparticles. In this regards, several methods for nanoparticle-protein interaction were developed and analysed in this review.Further, this paper reviews the current scientific development in nanoparticle-protein interactions.

Keyword:Nanoparticles (NPs), Goldnanoparticles (GNPs), Nanomedicine, Protein corona

1. Nanoscience and Nanomedicine

Nanoscience technology had outstanding developments in the past decades that enables usage of micro-scale and nano-scale materials in different areas of technology and medicine. Nanoscience depends on exact organization of nanoparticles in the order to get unique functionality [1].Nanoscientists aim to build new materials with ultimate properties, miniaturize existing products and of course provide us with deeper understanding of nature and life [2] . Nanoparticles, especially due to their size (100nm), have specific features that are ideal for manipulating biological interactions. In this regards, it is very important to understand nanoparticle interaction at cellular, sub-cellular and bio molecule level [3,4,5].Nanoparticles can be used as drug or drug carrier, acting like double-edged sword, they can be toxic agents or platform for therapy. Scientists are working hard to find nanoparticles that have most suitable properties for therapeutic applications [6], mostly for purposes of transporting small molecules as well as bio-macro - molecules to diseased cells or tissues [7].Furthermore, nanoparticles are being used to probe biological processes that are critical for diagnostics, but still there is no enough knowledge about potential risks from nanoparticles therapeutics applications [8].

Nanomedicine, based on application of nanoparticles in medicine has great capacity to treat viral or genetic diseases and even cancer, since smaller objects are more practical for the cell manipulation and disease treatment in humans. The national Cancer Institute and National Aeronautics and Space Administration, USA is working to develop nano sized technologies that can detect, diagnose and treat disease.Nanoscience and Nanomedicine could lead to next generation of products, diagnostics, and medical device, and enhanced gene therapy, tissue engineering procedures, medicine and medicine delivery techniques. On the other hand, knowledge about bio-compatibility and risk of exposure to nanoparticles is limited and there is a great need of comprehension the molecular mechanisms of interactions between nanoparticles and biological systems [6,9,10].

Understanding the protein-nanoparticles interactions is essential to stabilize and deliver protein-based therapeutic drugs and vaccines. In this regards, the purpose of this paper is to review the importance of nanoparticles-proteins interaction and to explain the basics of most important analytical methods responsible for characterization of nanoparticle-protein interaction. Furthermore, this review will focus on interaction of nanoparticles with bio molecules, with special emphasis on interaction of gold nanoparticles and proteins.

2. Nanoparticlesand Protein corona

Size of nanoparticles makes them able to enter in almost all parts of human body, including cells and organelles, while flat surfaces can only affect biological processes via cell surface receptors such as integrins[9, 11].When nanoparticles enter biological fluid, the first molecule that will react with nanoparticles, are proteins in more than 95% of all cases. The result of protein coating on nanoparticles surface is protein corona [10]. Protein corona may influence cellular uptake, inflammation, accumulation, degradation and clearance of nanoparticles[12]. Proteins of protein corona can change their native conformation, influencing the downstream regulation of protein-protein interactions, cellular signal transduction and transcription of DNA. For a better understanding of interactions between nanoparticles and proteins, we acquire information on binding affinities and stoichioetries for different combinations of proteins and nanoparticles. Adsorption of protein on nanoparticle surface is aided by hydrogen bonds, solvation forces, Van der Waals interactions, etc. [13,14]. Since different nanoparticles have distinct properties, the composition of protein corona is unique to each kind of nanoparticles and depends on many parameters [11].

The proteins corona can be hard and soft. It is thought that hard corona proteins interact directly with nanoparticle surface with high affinity, while soft corona consists of loosely bound proteins that interact with hard corona via weak-protein interactions [11]. Scientists proposed this model for protein corona as primary binders and soft corona as secondary binders that may influence hard corona and even prevent their interaction with the surrounding environment [15].It was shown that typical lifetime of hard corona can be even eight hours which indicates that hard corona defines the biological identity of the particle. Hard corona is formed in seconds, while the time of forming soft corona varies from seconds to hours. In contrast to hard corona with lifetime of around eight hours, soft corona will probably be desorbed in ten minutes [16]. Since there are many proteins that can absorb on nanoparticles in their biological environment, it is concluded that protein corona is not a fix layer and the composition of protein corona can be determined by kinetic rate of adsorption and desorption of each protein. Dynamic of protein corona as „biological identity" might lead to more clear classification system of nano-safety and could be used for engineering of nanomedical products [9].

2.1. Parameters affecting protein corona

Due to diverse parameters that affect protein corona the scientific community can understand why some nanomaterial's select specific type of proteins [11]. The most important parameters are explained in the following subparagraphs.

2.1.1. Surface charge of nanoparticle

Acting as important parameter in protein interaction, surface charge of nanoparticle can also denaturate the adsorbed proteins. It was found that proteins can denaturate when they interact with positively or negatively charged ligands, whereas neutral ligands can keep the native structure of proteins. In most study cases, positively charged nanoparticles were attracting proteins with isoelectric points less than 5, 5 such as albumin, but proteins with pI higher than 5.5 adsorb negatively charged particles [11].

2.1.2. Surface functionalization and Coating

Since nanoparticles are travelling through different environment of an organism, they may get pre-coated with different proteins and that pre-coatingcan determine which new proteins will bind to nanoparticle-protein complexes[10].

2.1.3. Hydrophobicity and Hydrophilicity

Nanoparticles withcharged orhydrophobic surfaces, tend to adsorb and denaturate more proteins than neutral and hydrophilic surfaces. This action happens due to clustering of hydrophobic polymer chain that forms distinct protein-binding sites [11]. Different particle features determine protein adsorption into nanoparticles, but hydrophobic interactions tend to be dominating feature [Schaefer, 2012]. In one study done in 2007, the affinity to nanoparticles of proteins that have uniform surfaces, increases with the increasing charge density and hydrophobicity [17].

2.1.4. Nanoparticle size

Same nanoparticles with different sizes have unlike compositions of protein corona[11], where nanoparticles have different sizes, from 70-700nmand that the amount of bound plasma proteins increased with increasing available surface area at constant particle weight[13].Nanoparticles surface area, available for protein binding, increases with decreasing of particle size [18].

2.1.5. Biological environment

Many in vitro studies have been done to examine nanoparticle-protein interaction, but it was still hard to predict the behavior of nanoparticles in living biological environment. That is why scientists are trying to use different cellular media in their experiments for better prediction in vivo events. In

most case   two different cellular media are used, Dulbecco's Modified Eagle's Media (DMEM) and Roswell Park Memorial Institute medium (RPMI) with differently sized citrate coated gold nanoparticles[19].

In the case of Gold nanoparticles, it is characterized that the dynamics of protein-nanoparticle interaction, can be differently mediated by different composition of cellular media. For example, DMEM cellular media induced a more abundant and stable protein corona on different sizes of gold nanoparticles, comparing to another cellular media,as withRPMI media [20].

3.   Gold Nanoparticles (GNP)

In the last several years Gold nanoparticles (GNPs) became main topic of interest for researches around the world. Interactions of inorganic particles with biomolecules, like proteins, are central field of interest to biotechnology and nanotechnology[21].GNPs have potential to be used for treatment of AIDS, tumors and Parkinson`s disease [22]. These nanoparticles have therapeutic benefits in treating cancer. It is well known that ionizing radiation does not make difference between malignant and normal cells, so healthy cells are also being damaged. When cancer targeting molecules, such as antibody, nucleic acid aptamer, oligopeptide and small targeting molecules are attached to the surface of GNPs cancer cell targeting is enhanced [4]. These appealing properties make GNPs good model particle system and standard platform for nanomedicine applications [6].

The attractive physical and chemical properties of GNP, they are most commonly used as inorganic nanoparticles for biological applications [8]. Gold nanoparticles in nanoscale size have very different physical and chemical properties than those of bulk metal [6].GNPs have been used for electron microscopy (EM) because of their size, density and electrical properties, which are ideal for accurate diagnostic properties, by enhancing the contrast of images and for electrochemical biosensors [23].

3.1. How nanoparticles enters body

Using NPs for drug delivery, imaging, labelling and visualizing applications require knowing the routes of NPs entrance in the body [2, 8], and how they affect the organism when they enter [8]. Primary interface for NPs with human body are respiratory system, gastrointestinal tract and skin. One of the most common ways of entering the NPs into the organisms is through the respiratory system. NPs less than 2.5 μm in size can pass easily, but larger particles are removed from the respiratory tract by the mucociliary escalator. NPs can also be eliminated by NP translocation into the

lymphatic system via their uptake by macrophages or NPs dissolution followed by the transfer of the products into the blood. Alveolar macrophages are also involved in the elimination of inhaled NPs. Interaction of macrophages with high concentrations of NPs can cause asthma, obstructive pulmonary disease or respiratory infection.In the case of moderate concentrations of NPs, it may cause autophagy or apoptosis of lung cancer, which is actually positive effect [11]. Skin serves as strong barrier for everything that can enter human body. But, certain types of NPs use skin for entering body, for example TiO_2 particles found in cosmetics. Skin wounds can make entrance of NPs easier. And the third major way of NPs entrance into the body is gastrointestinal tract (GIT). When NPs enter the GIT they have two possibilities, entering the bloodstream, and soon getting to the liver or lymphatic system that may cause immune response [2].NPs can also be injected directly into the bloodstream where competition of different biomolecules can occur in order to be adsorbing on the surface of the NPs. At the beginning, at so called early stage, most abundant proteins (mainly albumin, IgG, fibrinogen and apolipoprotein) will be the first ones to adsorb, but later on those proteins will be replaced by proteins of lower abundance, higher affinity, and slower kinetics[2]. In this regards, other immunoglobulins, apolipoproteins and components of the complement system can later be nanoparticle-bound proteins. This phenomenon of competitive adsorption of proteins onto a surface based on protein concentrations, affinities and incubation time is called Vroman`s effect [11,24]. Vroman`s effect is very important in determining how the protein corona forms as the nanoparticle is moved from one part of the body to another [24].

4.   Protein Adsorption on Nanoparticle Surface

After inhalation or ingestion very small portion of NPs (1% or less) will move to the bloodstream. There is a great chance that NPs agglomeration will occur too. Bio-distribution of NPs is very important and depends on adsorption of proteins in the different body parts and liquids. A model for characterization of NPs- small organic molecules biological surface adsorption index approach (BSAI), was developed and discussed. The model can applied for characterization of interactions between NPs and biomolecules (proteins, amino acids etc.) [10,25, 26]. NP-protein interactions are changing all the time, even within the same environment. Different factors can affect the kinetics of protein adsorption on the NP surface. One of the factors that influence the nanoparticle protein corona (NP-PC) composition is amount of proteins that may interact with NP surface. Apart from the amount of proteins, affinity of the protein toward the NP surface also affects the adsorption. Different proteins can arrange

themselves differently on the NP surface. For example, plasma proteins such as human serum albumin adsorb in a monolayer on iron-platinum NP surface. The way the proteins arrange in this situation, can affect the biological reactivity of the complex (10).

5. Effects on Protein Conformation

Proteins are complex molecules made from long chains of amino acids. Each protein type differs in the sequence which makes the protein unique. That sequence determines protein`s shape, structure, function, and protein`s three-dimensional shape which is the native conformation of the protein[27]. The native conformation is stable, but can be easily disrupted by the interactions with the surface. The function of a protein changes with changing the conformation [9].

Proteins are known as surface active molecule; they can be adsorbed and accumulated in the interactions with other molecules which can even result in protein loss and degradation [28]. Proteins that are adsorbed on NPs can be used as biosensors and in drug delivery. This is why our understanding of the effect of NPs on the protein structure is crucial for nanomedicine application [14]. When proteins bind to planar surface, it is very possible that the changes in secondary structure will occur. The curved shape of NP surface can help proteins to keep their native conformation [30].

Nevertheless, different studies on NP –protein interaction indicate that disturbance of protein structure will appear in many cases [11]. By using fluorescence spectroscopy, we can observe the misfolded proteins that lose their normal biological activity [6]. When proteins adsorb on the surface of NPs, nanoparticle protein corona formation occurs depending on the characteristics of both NP and protein, and local environment. NP surface can change the structure of the adsorbed protein thus affecting the overall bio-reactivity of the NP [2,10]. For example, Gold nanoparticles can change structure in the bovine serum albumin (BSA) but no conformational changes occur when BSA adsorb to carbon C60 fullerene NP [10]. It is discovered that some proteins can keep their native structures, like RNAse and lysozyme on silica NP[29]. Hovewerm, a differenet study discovered also irreversible structural changes in albumin and lactoperoxidase on silica NP, proving that conformational changes depend on type of the protein and nanoparticles[10]. For biological applications of nanoparticle protein corona, it is very important that labelling does not change the protein structure so the protein can stay functional and active. Its is shown that binding and dissociation parameters of protein-NP complex depend on the surface characteristic of NP as well as physiochemical properties of the protein [1].

During the stronger binding of the proteins to the NPs, it is natural that conformational changes occur.

All the parameters that affect the strength of protein-NP binding also affect relative value of the protein-NP binding equilibrium constant (K). Protein-NP binding constant (K) quantifies the relative strength of the protein-NP binding [6]. Equilibrium constant for plasma proteins and GNPs are in the range of 104 to 107 (mol/L-1).

In the study on interaction of GNPs with human blood proteins it is observed increase of binding constant where NP size was in the range between 5 and 60 nm. This is in contrary to findings that show that binding association constant can decrease when NPs size is larger than about 80 nm [6].

There are thousands of different proteins in real body fluids, and those proteins actually compete to adsorb on the NPs surface. Change in conformation lead to peptide aggregation and the formation of amyloid fibrils. This was used for tracking development of several neurodegenerative diseases. Scientist shows that some NPs can inhibit the fibrillation of the disease-associated amyloid β protein [14].It can be concluded that still there is no enough information about how NP adsorption induces the level of protein conformational changes, even under conditions where protein binding constants are rather similar [6].

6. Nanoparticle role inProtein Fibrillation

Protein fibrillation is defined as process by which misfolded proteins form large linear aggregates or amyloid fibrils [31]. Nanoparticles have great potential to enhance rate of protein fibrillation. Their potential to induce protein fibrillation is a function of both the NP surface charge, which promotes adherence of the protein, and its large surface area. The observation of fibrillation, which is a specific kind of aggregation phenomenon relevant for amyloid proteins, raises the possibility that NPs could play a role in increased risk of amyloidosis and other protein-misfolding diseases [32].Depending on their physiochemical properties, nanoparticles can have various effects on kinetics of amyloid beta peptide fibrillation process. For instance, positively charged super-magnetic iron oxide nanoparticles are capable of promoting fibrillation compared to uncharged nanoparticles at same particle concentration [33]. In addition, with their enormous surface area which act as a scaffold for protein adsorptions they offer significant potential for probing the mechanisms of protein fibrillation, and in the longer term for diagnosis or even treatment of amyloidogenic diseases [32,33].

Numerous proteins and peptides are responsible for making amyloid fibrils during specific disease formation, including amyloid beta peptide, glucagon, polyglutamine protein, but most commonly used is amyloid beta peptide. Since amyloid beta peptide is amphipathic molecule that is prone to self-association and formation of fibrils, it is used as model peptide to investigate the effects of nanoparticles on fibrillogenensis [34].

7. Methods Used For Protein-Nanoparticle Interactions

Proteins that are part of nanoparticle protein coronaare being isolated and identified in order to understand bio-reactivity of specific NPs. Methods that can be used to study NP-protein interactions are numerous, but the most common are Fourier transform infrared spectroscopy (FTIR), Circular Dichroism (CD) spectroscopy, Isothermal titration calorimetry (ITC), Surface plasmon resonance (SPR), Fluorescence spectroscopy, Nuclear magnetic resonance (NMR) spectroscopy but the most used method for studying nanoparticle protein corona is Mass spectrometry (MS). Mass spectrometry can be used for many types of NPs and is used to identify proteins and quantify amounts of proteins adsorbed on NP surface [9,10]. Combination of these methods should provide more complete picture about protein-nanoparticle interactions; protein structure and conformation upon their interaction with nanoparticles [6].

7.1. Fourier transform infrared spectroscopy (FTIR)

FTIR provides useful information on protein-nanoparticle interaction. FTIR has ability to give detailed description of bonds, their lengths, strengths, angles and conformation, chemical and physical structure, redox state, hydrogen bonding, electric fields, conformational freedom and dynamic motions [28]. The goal of any absorption spectroscopy is to measure how well a sample absorbs light at each wavelength, so that this technique.

7.2. Raman spectroscopy

It is another important method for studying protein-nanoparticle interactions. Raman spectroscopy involves interaction of light from the laser and a molecule, which results in the scattered light. After that interaction, light is scattered from the sample and the frequency shifts according to the frequency of the molecular vibrational mode [28].

Raman spectroscopy and FTIR are not so often used for studying the electrostatic interactions even though these two methods can provide important information because of the characteristics bands that many charged residues exhibit in the vibrational spectrum [28].

7.3. Circular dichroism (CD) spectroscopy

Along with the FTIR and Raman spectroscopy, CD spectroscopy is valuable method for determining protein secondary structure. It is valuable tool to study the interaction of proteins with other molecules too. CD spectroscopy involves absorption in the far (180-250 nm) and near (250 nm-visible) UV spectrum. It provides information about protein secondary and tertiary structure upon binding [6,28]. One of the most commonly used technique for analyzing secondary structure is based on far-UV CD spectroscopy [14].

7.4. Fluorescence spectroscopy

Fluorescence spectroscopy can be used for studying proteins, determining protein structure, size, shape and dynamics. It is proven to be very fruitful in the studies of the protein`s structure and conformation [6]. In its base, this method uses a beam of light that excites the electrons in the molecules of certain compounds, which makes them emitting light [28].

One of the factors that regulates the spectroscopy accuracy is Photoluminescence (PL) quenching, widely used because of its sensitivity, reproducibility and suitability. In general, decrease of fluorescence intensity by interaction of the excited state of the fluorophore with its surroundings is known as quenching. For example, Gold can efficiently quench the emission of many chromophores. Chromophores are molecules that adsorb light [6]. Furher, the emission typical for tryptophan, tyrosine and phyenyline residues in proteins are used for investigating binding and conformation changes of proteins when interacting with small molecules like nanoparticles. When protein binds to the NP, tryptophan, tyrosine and phyenyline residues become accessible to the metallic surface of the NP are then quenched. By decreasing the NP size, the surface area is increases, making smaller NPs more efficient fluorescence quenchers than larger ones.Conformational changes of a protein can, in principle, be evaluated by the measurement of changes in the peak intensity wavelength λmax in the continuous fluorescence emission intensity spectrum [6].

7.5. Nuclear magnetic resonance spectroscopy (NMR)

NMR spectroscopy provides useful information on protein structure and it was used for studying interactions

of ubiquitin molecules with gold nanoparticle surface. For example, the information obtained with NMR spectroscopy provided possibility to distinguish the exact ubiquitin domain that bound to the NP surface [10]. In general, this methods is a powerful technique, which directly analyses the molecularscale of nanoparticle formation and morphology in situ, in both the solid and the solition phase, for the study of noble metal nanoparticles[35].

7.6. Isothemal titration calorimetry (ITC)

Isothermal titration calorimetry (ITC) is used to directly measure the thermodynamics of protein-ligand interactions [28]. In combination with other techniques can contribute to more complete understanding of NPs-biomoleculesinteraction. All these techniques provide data quickly and require minimal sample preparation [28].

8. Conclusion

The nanoparticle protein interaction represents the basis of bioreactivity of nanoparticles. These interactions results in the formation of a dynamic protein corona around nanoparticles. The protein corona is responsible for various physiologicalfunctions such as absorption. In

addition,its is shown that nanoparticle surface is important factor resulting in conformational changes in the adsorbed protein molecules and, therefore, affect the overall nanoparticle bioreactivity.Due to their beneficial characteristics' gold nanoparticles (GMPs) are the most commonly used inorganic nanoparticles for biological applications. When GNPs enter a biological fluid like blood, proteins adsorb on the surface of GNPs, and GNP-protein corona formation occurs depending on the characteristics of both protein (protein corona) and local environment.A significant increase in biomedical applications of nanomaterials and their potential toxicity requires deverse analytical techniques to determine protein – nanoparticle (NP) interactions. Most important tecnhiques for the analysis of binding affinity, binding ratio, and binding mechanisms of NP-protein interaction, are based on spectroscopy. Since NP-protein binding is a dynamic event, Fourier transform infrared spectroscopy (FTIR), Nuclear magnetic resonance spectroscopy (NMR), Fluorescence spectroscopy and other methods are used. Understanding of the effect of NPs on the protein structure is crucial for nanomedicine application. Further studies on the formation and the composition of the protein corona needs to be done, including all the possible consequences that are incurred by the specific proteins.

References

[1] Singh, V., Nair, S. P., & Aradhyam, G. K, Chemistry of conjugation to gold nanoparticles affects G-protein activity differently. *J Nanobiotechnology*,11(7) ,2013.

[2] Shemetov, A. A., Nabiev, I., & Sukhanova, A, Molecular interaction of proteins and peptides with nanoparticles. *ACS nano*, *6*(6), 4585-4602,2012.

[3] Qingxin Mu, Guibin Jiang, Lingxin Chen, Hongyu Zhou, Denis Fourches, Alexander Tropsha,and Bing Yan, Chemical Basis of Interactions Between Engineered Nanoparticles and Biological Systems, *Chem Re,* 13; 114(15): 7740–7781, 2014.

[4] Li, H., LaBean, T. H., & Leong, K. W, Nucleic acid-based nanoengineering: novel structures for biomedical applications. *Interface Focus,* rsfs20110040, 2011.

[5] Tapan K. Jain, Marco A. Morales, Sanjeeb K. Sahoo, Diandra L. Leslie-Pelecky, VinodLabhasetwar. Iron Oxide Nanoparticles for Sustained Delivery of Anticancer Agents, *Molecular Pharmaceutics,* 2 (3), 194 -205, 2005.

[6] Lacerda, S. H. D. P., Park, J. J., Meuse, C., Pristinski, D., Becker, M. L., Karim, A., & Douglas, J. F, Interaction of gold nanoparticles with common human blood proteins. *ACS nano*, *4*(1), 365-379, 2009.

[7]Rochelle Arvizo, Resham Bhattacharya, and Priyabrata Mukherjee. Gold nanoparticles: Opportunities and Challenges in Nanomedicine, *Expert Opin Drug Deliv,* 7(6): 753–763.2010.

[8] Yang, Y., &Burkhard, P. Encapsulation of gold nanoparticles into self-assembling protein nanoparticles. *J Nanobiotech*, *10*, 42, 2012.

[9] Lynch I and Dawson KA. Protein-nanoparticle interactions, *Nano Today*, 3:40–47, 2008.

[10] Saptarshi, S. R., Duschl, A., &Lopata, A. L.Interaction of nanoparticles with proteins: relation to bio-reactivity of the nanoparticle. *J Nanobiotechnol*, *11*(1), 26,2013.

[11] Rahman, M., Laurent, S., Tawil, N., Yahia, L. H., &Mahmoudi, M. Nanoparticle and protein corona. In *Protein-nanoparticle interactions,Springer Berlin Heidelberg,* 21-44, 2013.

[12] Zhu, Z. J., Posati, T., Moyano, D. F., Tang, R., Yan, B., Vachet, R. W., &Rotello, V. M. The interplay of monolayer structure and serum protein interactions on the cellular uptake of gold nanoparticles. *Small*, *8*(17), 2659-2663, 2012.

[13] Lynch I, Lundqvist M, Stigler J, Elia G, Cedervall T, Dawson KA. Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts. Proc Natl Acad Sci USA, 105:14265–14270, 2008.

[14] Chatterjee, T., Chakraborti, S., Joshi, P., Singh, S. P., Gupta, V., &Chakrabarti, P. The effect of zinc oxide nanoparticles on the structure of the periplasmic domain of the Vibrio choleraeToxR protein. *Febs Journal*, *277*(20), 4184-4194, 2012.

[15] Chao Y, Karmali PP, Simberg. Role of carbohydrate receptors in the macrophage uptake of

dextran-coated iron oxide nanoparticles. *AdvExp Med Biol* 733:115–123, 2012.

[16] Monopoli MP, Walczyk D, Campbell A, Elia G, Lynch I, Bombelli FB, Dawson KA. Physical-chemical aspects of protein corona: relevance to in vitro and in vivo biological impacts of nanoparticles. *J Am Chem Soc*, 133:2525–2534, 2011.

[17] De M1, You CC, Srivastava S, Rotello VM. Biomimetic interactions of proteins with functionalized nanoparticles: a thermodynamic study. *J Am Chem Soc.*129(35):10747-53, 2007.

[18] Lundqvist M, Stigler J, Cedervall T, Berggard T, Flanagan MB, Lynch I, Elia G, Dawson K. The evolution of the protein corona around nanoparticles: a test study. *ACS Nano*, 5:7503–7509, 2011.

[19] Maiorano G, Sabella S, Sorce B, Brunetti V, Malvindi MA, Cingolani R, Pompa PP. Effects of cell culture media on the dynamic formation of protein-nanoparticle complexes and influence on the cellular response. *ACS Nano,* 4:7481–7491, 2010.

[20] Sabuncu AC, Grubbs J, Qian S, Abdel-Fattah TM, Stacey MW, Beskok A. Probing nanoparticle interactions in cell culture media. *Colloids Surf B Biointerfaces*, 15;95:96-102, 2012.

[21] Hoefling, M., Monti, S., Corni, S., & Gottschalk, K. E. Interaction of β-sheet folds with a gold surface. *PloS one*, 6(6), e20925.2011.

[22] Ye, M., Tang, L., Luo, M., Zhou, J., Guo, B., Liu, Y., & Chen, B. Size-and time-dependent alteration in metabolic activities of human hepatic cytochrome P450 isozymes by gold nanoparticles via microsomal co-incubations. *Nanoscale research letters*, 9(1), 1-16, 2014.

[23] Yang, M., Kostov, Y., Bruck, H. A., &Rasooly, A.Gold nanoparticle-based enhanced chemiluminescenceimmune sensor for detection of Staphylococcal Enterotoxin B (SEB) in food. *International journal of food microbiology*, 133(3), 265-271, 2009.

[24] Aggarwal, P., Hall, J. B., McLeland, C. B., Dobrovolskaia, M. A., & McNeil, S. E, Nanoparticle interaction with plasma proteins as it relates to particle biodistribution, biocompatibility and therapeutic efficacy. *Advanced drug delivery reviews*, 61(6), 428-437, 2009.

[25] Schaefer, J., Schulze, C., Marxer, E. E. J., Schaefer, U. F., Wohlleben, W.,Bakowsky, U., & Lehr,

.

C. M. Atomic force microscopy and analytical ultracentrifugation for probing nanomaterial protein interactions. *ACS nano*, 6(6), 4603-4614, 2012.

[26] Omanovic-Miklicanin, E., Valzacchi, S., Simoneau, C. Gilliland, D., & Rossi, F.Solid-phase micro extraction/gas chromatography–mass spectrometry method optimization for characterization of surface adsorption forces of nanoparticles. *Analytical and bio-analytical chemistry*, 406(26), 6629-6636, 2014.

[27] Albert's, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P.Molecular Biology of the Cell. New York: *Garland Science*; *Classic textbook now in its 5th Edition*, 2002.

[28] Kamerzell, T. J., Esfandiary, R., Joshi, S. B., Middaugh, C. R., &Volkin, D. B. Protein–excipient interactions: Mechanisms and biophysical characterization applied to protein formulation development. *Advanced drug delivery reviews*, 63(13), 1118-1159, 2011.

[29] Turci F, Ghibaudi E, Colonna M, Boscolo B, Fenoglio I, Fubini B. An Integrated Approach to the Study of the Interaction between Proteins and Nanoparticles. *Langmuir*. 26:8336–8346, 2009.

[30] Walkey CD, Chan WC. Understanding and controlling the interaction of nanomaterials with proteins in a physiological environment. *Chem Soc Rev* 41:2780–2799, 2012.

[31] Q A Pankhurst, NTK Thanh, SKJones and J Dobson. Progress in applications of magnetic nanoparticles in biomedicine. *Phys. D: Appl. Phys.* 42, 224001 (15pp), 2009.

[32] Nanoparticles as catalysts for protein fibrillation, National Academy of Science. *Proc Natl Acad Sci U S A.* 22; 104(21): 8679–8680, 2007.

[33] Mahmoudi M, Kalhor HR, Laurent S, Lynch I Protein fibrillation and nanoparticle interactions: opportunities and challenges.*Nanoscale.* 7;5(7):2570-88, 2013.

[34] Mahmoudi M, Monopoli MP, Rezaei M, Lynch I, Bertoli F, McManus JJ, Dawson KA. The protein corona mediates the impact of nanomaterials and slows amyloid beta fibrillation.*Chembiochem*.18;14(5):568-72, 2013.

[35] Lauren E. Marbella and Jill E. Millston**.** NMR Techniques for Noble Metal Nanoparticles. *Chem. Mater.* 27 (8), 2721–2739, 2015.

An in silico approach for structural and functional analysis of Heavy Metal Associated (HMA) proteins in *Brassica oleracea*

Jasmin Šutković	Mujo Kekić	Maida Ljubijankić	Peter Glamočlija
International University of Sarajevo,	International University of Sarajevo,	International University of Sarajevo,	International University of Sarajevo,
Hrasnicka cesta 15, Sarajevo,	Hrasnicka cesta 15, Ilidza,	Hrasnicka cesta 15, Ilidza,	Hrasnicka cesta 15, Ilidza,
Bosnia and Herzegovina	Bosnia and Herzegovina	Bosnia and Herzegovina	Bosnia and Herzegovina
jsutkovic@ius.edu.ba	mujo_kekic@hotmail.com	maidaaa-k@hotmail.com	ammar_glamoclija@live.com

Abstract

Heavy metal ATPases (HMAs) are the most important proteins involved in heavy metal accumulation process. Brassica oleracea has 5 HMA (1-5) homologues whose 3D structure has been predicted and validated in this study by different bioinformatics tools. Phylogenetic and multiple sequence alignment analyses showed high relationship between HMA2 and HMA4, while two same domains were identified in all five HMA proteins: E1-E2 ATPase and haloacid dehydrogenase (HAD) domain. Four HMA (2-5) proteins were identified to be localized in the plasma membrane, while HMA1 localization is predicted to be in plastid. Interactome analysis revealed high interaction of all HMA (1-5) proteins with many metal ion binding proteins and chaperones. Among these, interesting and strong interaction is observed between all HMA (1-5) proteins and ATX1, while HMA1, HMA2 and HMA4 have been found to strongly interact with FP3 (farnesylated protein 3) and FP6 (farnesylated protein 6) proteins. Docking site predictions and electrostatic potentials between HMA2/HMA4 and the interactome proteins were explained and discussed in this study.

Keywords: Protein structure prediction, Heavy metals; accumulation; transport; interactome; docking site

1. Introduction

1.1. Brassicaceae family

Brassicaceae is the name of a medium-sized and economically important family of flowering plants, informally known as the mustards, mustard flowers, the crucifers, or the cabbage family. The name is derived from the included genus Brassica. Brassica is a one of the major crop worldwide with *Brassica oleracea* as a main consumed species in Europe and USA. Brassica is a genus with many beneficial characteristics for our health, such as reducing risk for age related chronical illness, degenerative diseases and it reduces risk of several types of cancer. Brassica contain many vitamins which are essential for our health, such as vitamin A, C, E, K and B-6, carotenoids (such as c- and b-carotene and zeaxanthin), anthocyanins, folate, soluble sugars and phenolic compounds which are known to be the major antioxidants of Brassica crops [1].

Interesting fact is that all parts of Brassica is used as a food, including root, stems, leaves, flowers, buds and seeds. Brassica has many species, thanks to difference in phenotype within themselves. Like all species in Brassica family, *Brassica oleracea* is very rich with vitamins and other nutrients. *Brassica oleracea* has been bred into a wide range of cultivars, including cabbage, broccoli, cauliflower, brussels sprouts, collards, and kale, some of which are hardly recognizable as being members of the same genus, let alone species [2]. Brassica vegetables are highly regarded for their nutritional value. With high amounts of vitamin C and soluble fiber they are excellent candidates to fight cancer, including molecules known of anticancer properties such as cellsproperties:3,3'-diindolylmethane, sulforaphane and selenium [3]. Furthermore, Brassica vegetables are rich in indole-3-carbinol, a chemical which boosts DNA repair in cells in vitro and appears to block the growth of cancer cells in vitro [4]. They are also a good source of carotenoids, with broccoli having especially high levels [5] and goitrogens, some of which suppress thyroid function [6].

1.2. Brassica oleracea genetic characterization

A recent study done with AFLP markers, evaluated the genetic diversity in kale landraces through Europe and compared the diversity to that in the wild populations of *Brassica oleracea*. In total 17 accessions were collected from all around Europe, including Bosnia and Herzegovina, Croatia and Turkey. In Bosnia and Herzegovina 47 individuals were analyzed, in Rivine, Dubrave and city of Stolac and its interesting to say that among a total of 93 polymorphic markers which were scored, a unique allele was found in only one accession, and it is the one in Bosnia and Herzegovina. In addition, the AFLP analyses of genetic diversity in leafy kale (*Brassica oleracea L. convar. acephala*) landraces, showed that Herzegovina has a 58% of polymorphic loci, while Croatia had 69% and Turkey 76%. Accessions from Bosnia and Herzegovina, Croatia, Portugal and Turkey contain many individuals with mixed genotype, sharing parts of their genome with other accessions due to common ancestry or gene flow [7].

1.3. Brassicaceae family as heavy metal accumulators

Through various ways, as for example, gas exhausts, energy and fuel production, intensive agriculture, and sludge dumping activities, humans contaminate soils and aqueous streams with large quantities of toxic metals. A number of studies from developing countries have reported heavy metals contamination in wastewater and wastewater irrigated soils [8]. In this regards, heavy metals are harmful to humans and other life forms, as they can cause cancer, blindness, loss organ function, severe illness, and death. The fact that some of these Brassicaceae family plants can accumulate high amounts of toxic metals, without visible symptoms, and in the same time being important food crops as well, leads to potential contamination of our food chain and this has to be taken into account in any phytoremediation process [9]. In general, plants require at least 14 mineral elements for their nutrition. These include the macronutrients nitrogen (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg) and sulphur (S) and the micronutrients boron (B), chlorine (Cl), iron (Fe), manganese (Mn), copper (Cu), zinc (Zn), nickel (Ni) and molybdenum (Mo). Crop production is often limited by low bioavailability of essential mineral elements and/or the presence of excessive concentrations of potentially toxic heavy metals, such as Fe, Mn, Cu, Cr, Cd, Pb, Zn and Al in the soil solution. High concentrations of heavy metals in the soil can inhibit plant growth and reduce crop yields, which can affect sustainable development severely [10]. Therefore, some known Brassicaceae family species are already proven to be effective heavy metal accumulators. For example, *A.halleri* is one of the best model organism for the study of plant adaptation to extreme metallic conditions since it is considered as Zn and Cd hyper-accumulator [11]. *A.halleri* is found to cope with excessive metal ions and toxicity in a way that it uses effective metal uptake, increased xylem loading and increased detoxification in shoot tissues. In recent years, several types of transporters involved in these processes have been identified in Zn and Cd hyper-accumulators, specifically in *A.halleri* [12]. The most investigated proteins that transfer the toxic metals are named as Heavy metal ATPases. These are located within membrane complexes of plant cell. As their name implies they produce or utilize energy in form of ATP. There are three different types of heavy metal transporters: P-type, V type and F0-F1 type. Most common type that is found in plant organisms is P-type. The proteins of this type usually transport essential metal ions which are Cu^{2+}, Zn^{2+}, Mn^{2+}, Fe^{2+} and Co^{2+}. This type of transporters does not produce energy, but actually uses it in order to pump these metals. Another type is V type which also utilizes energy. The third type is F0-F1 ATPases which produce energy instead of using it. The function of these proteins is to regulate the concentration of these metals in all tissues found in plants [10]. Their proper functioning is highly important for plant, where high levels of essential and some non-essential metals can be very toxic for the plant [9]. For example, expression of the HMA1 gene from *Atriplex canescens* significantly increased the ability of yeast cells to adapt to and recover from exposure to excess iron. AcHMA1 expression also provided salt, alkaline, osmotic and oxidant stress tolerance in yeast cells. In this regards these results suggest that HMA1 gene encodes a membrane-localized metal tolerance protein that mediates the detoxification of iron in eukaryotes and may be involved in the response to abiotic stress [13]. HMA2 is known for maintaining plant metal homeostasis by transporting Zn and Cd metal ions [14]. It is shown that HMA2 and HMA4 drive metal efflux out of the cell in *A. thaliana* [15] and promote xylem loading of metal in *N. caeruslecens* [16]. HMA4 is responsible for zinc hyper-accumulation in *A. halleri* as it shown by a RNA interference approach for down regulation of its expression. Additionally, transfer of the HMA4 gene to *A. thaliana* enables zinc partitioning into xylem vessels and up-regulated specific genes characteristic for zinc hyper accumulators [17]. This example shows impressively the importance of regulatory gene expression and gene copy number expansions for the special trait of metal hyper-accumulation. Furthermore, AtHMA4 is shown to be responsible for the reduction of Cd uptake/accumulation [18]. In contrast, HMA3 is localized at the tonoplast enabling vacuolar metal influx and therefore cellular sequestration [19]. The

quantitative trait locus (QTL) analysis on chromosome 1 in Arabidopsis thaliana revealed that this QTL regulates Cu translocation capacity and involves Cu-transporting via HMA5 [20].

Furthermore, in Arabidopsis, the heavy metal P-type ATPase HMA5 is shown to interact with metallo-chaperones and function in copper detoxification of roots [21]. It is found that some HMA genes are highly expressed in *A.halleri*, suggesting their importance in hyperaccumulation process. Among HMA genes, HMA2 and HMA4 are discovered to be among the most important ones with HMA4 playing role ineffective root-to-shoot Zn/Cd translocation [12] and HMA3 playing role in Zn detoxification [22]. Additionally, HMA2 and HMA4 have both been demonstrated to be plasma membrane proteins. Finally, both of these proteins appear to function in Cd transport within the plant. Analysis of whole plant demonstrates that HMA2 accumulate more Zn and Cd than wild type plants although they do not appear to have an increased sensitivity to either metal. HMA4 mutant plants accumulate more Zn and Cd in the roots but they accumulate less Zn and Cd in leaves [22].

2. Material and methods

2.1. Retrieving HMA sequences and Multiple Sequence Alignment

The sequences of Heavy metal ATP proteins were obtained from the National Center for Biotechnology Information (NCBI) Protein Database [23]. Sequences' accession numbers are listed in table 1.

Table 1: Accession numbers of HMA proteins

HMA proteins	*Arabidopsis thaliana*	*Brassica oleracea*	*Brassica napus*	*Brassica rapa*
HMA1	At4g37270	XP_013595810.1	XP_013722994.1	XP_009137459.1
HMA2	At4g30110	XP_013609995.1	XP_013748024.1	XP_009128090.1
HMA3	At4g30120	XP_013591300.1	XP_013704754.1	XP_009128077.1
HMA4	At2g19110	XP_013629797.1	XP_015765843.1	XP_009150707.1
HMA5	At1g63440	XP_013606061.1	XP_013684388.1	XP_009112946.1

Multiple sequence alignment (MSA) has been performed using the Clustal Omega software located on the website of the European Bioinformatics Institute (EBI), using default options [24]. Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences. MSA is an invaluable bioinformatics tool used to measure the similarity between sequences, examine patterns of conservation and variability and derive evolutionary relationships [25].

2.2. Phylogenetic tree construction

In order to infer the evolutionary relationship between the HMA proteins, a phylogenetic tree was constructed using ClustalW2 phylogeny, a web service for phylogenetic analysis of molecular sequences [26]. The service was run on default settings, and the steps that it performed to construct the phylogenetic tree involved MSA, alignment organization and construction, and visualization of the phylogenetic tree using different integrated tools.

2.3. 3D structure prediction and validation

The structures of HMA proteins were hereby predicted with the help of the Phyre2 protein homology modeling server [27]. Phyre2 is a web-based service for protein structure prediction that is free for non commercial use, and being one of the most popular methods for protein structure prediction. Cited over 1000 times, it is able to generate reliable protein models. Phyre2 has been designed and funded by the Biotechnology and Biological Sciences Research Council (BBSRC) from United Kingdom. A practical and widely cited molecular visualization tool PyMOL was used for structure visualization and representation. PyMOL vs 1.31 edu (The PyMOL Molecular Graphics System), is a molecular visualization tool that provides viewing, customizing and exporting of the visualized molecules.

The validation and stereo-chemical analysis of the predicted structures was performed using several tools. The first one was QMEAN6, available as the structure assessment tool at ExPASy server. QMEAN6 is a scoring function that is actually a linear combination of six terms: torsion angle potential over three consecutive amino acids, two distance-dependent interaction potentials, solvation potential and two terms describing the agreement of the predicted structure and the solvent accessibility of the model [28]. Also provided is the Z-score of the QMEAN6, which compares the estimated score to the score from a high-resolution reference structure solved experimentally by X-ray crystallography, with strongly negative Z-scores expected from low quality models.

In QMEAN6 score better predictions have higher scores (between 0 and 1) and in Z-score lower quality predictions have more negative scores. Next, we used Verify3D, which assesses protein structures using three-dimensional profiles, analyzing the compatibility of a 3D model with its own amino acid sequence (1D). The scores range from -1 (bad score) to +1 (good score) [29]. Stereo-chemical quality of the protein models was assessed with the PROCHECK software [30]. PROCHECK compares the geometry of the residues in the predicted model with the known stereo-chemical values from well-known structures. It results in Ramachandran plots, providing information about the dihedral angles φ and ψ of amino acid residues in the protein structure. What PROCHECK basically does is comparison of the geometry of the residues in the predicted model with the known stereo-chemical values from well-known structures [30].

2.4. Localization of proteins

For the subcellular localization, we used the recently developed tool PSI (Plant Subcellular localization integrative predictor) which uses the group voting strategy and machine learning to combine the results of 11 independent subcellular localization tools: cello, mPloc, Predotar, mitoProt, MultiLoc, TargetP, WolfPSORT, subcellPredict, iPsort, Yloc and PTS1 [31].

2.5. Domain search and interaction prediction

The identification of domains in the five HMA proteins was performed using the online tool SMART (Simple Modular Architecture Research Tool) located on the website of the European Molecular Biology Laboratory (EMBL). The tool is able to detect more than 500 domain families from chromatin-associated, extracellular and signaling proteins [32].

Interactome of the HMA proteins was determined by using the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins). It was used for searching interlogs of the five proteins. The STRING database consists of known and predicted protein interactions of currently 9 643 763 proteins and 2031 organisms. The predicted protein interactions are classified into physical or functional associations. What the program does basically is to determine binary interactions of each individual protein with predicted proteins.

Further, each interaction is assigned to a confidence score which depicts the quality and number of experimental technique used for the detection of these protein interactions [33].

2.6. Docking sites prediction

The docking site prediction was undertaken by the ClusPro 2.0 online software, an automated docking tool from the Structural Bioinformatics Lab of Boston University. ClusPro works on the basis of providing 70 000 rotations for the ligand protein from which 1000 rotations with the lowest score are chosen [34].

3. Results

3.1. MSA and Phylogenetic tree construction

Multiple sequence alignment was performed by the Clustal Omega online tool. The results of the sequence alignment are in table 2.

Table2. Multiple sequence alignment of Brassica oleracea HMA proteins

Seq. 1	Length in aa	Seq.e 2	Length in aa	Clustal Omega score in %
HMA 1	817	HMA 5	999	26.81
HMA 1	817	HMA 3	758	25.60
HMA 1	817	HMA 2	883	26.24
HMA 1	817	HMA 5	999	26.41
HMA 5	999	HMA 3	758	27. 82
HMA 5	999	HMA 2	883	29.94
HMA 5	999	HMA 4	1195	29.15
HMA 3	758	HMA 2	883	69.14
HMA 3	758	HMA 4	1195	67.60
HMA 2	883	HMA 4	1195	71.71

aa: Amino acids

Using the same tool we have constructed the phylogenetic tree that shows the evolutionary relationship between the aligned HMA proteins from *Brassica oleracea* .

Figure1. Phylogenetic tree (cladogram) of HMA proteins in Brassica oleracea

In figure 1 we observe that we have two sister groups, one group being HMA2 and HMA4 proteins and on another side HMA1 and HMA5 proteins, each group having a common ancestor. HMA3, as a lone taxon, shares common ancestor with HMA1 and HMA5 but is more distant. However, it shows more homology with HMA2 and HMA4, as confirmed by Table 2.
In addition, phylogenetic tree was constructed *combining the five B.oleracea* HMA proteins with

HMA proteins from *A. thaliana* and *B.rapa* (see Figure 2). In this figure, it's clearly visible that *A.thaliana* HMA proteins share the common ancestors with all other taxa analyzed in the phylogenetic tree. Furthermore, this indicated the possibility of an evolutional change of *A.thaliana* HMAs into *B.rapa* and *B.oleracea* HMA proteins. The HMA proteins from *B.rapa* and *B.oleracea* share the common ancestor, being sister taxa with all HMA proteins, except HMA3. HMA3 protein from *B.rapa* and *A.thaliana* shares the common ancestor whereas *B.Oleracea* has evolved separately.

HMA1A.THALIANA
HMA1B.RAPA
HMA1B.OLERACEA
HMA5A.THALIANA
HMA5B.RAPA
HMA5B.OLERACEA
HMA3A.THALIANA
HMA3B.RAPA
HMA3B.OLERACEA
HMA4B.OLERACEA
HMA4A.THALIANA
HMA2A.THALIANA
HMA2B.RAPA
HMA2B.OLERACEA
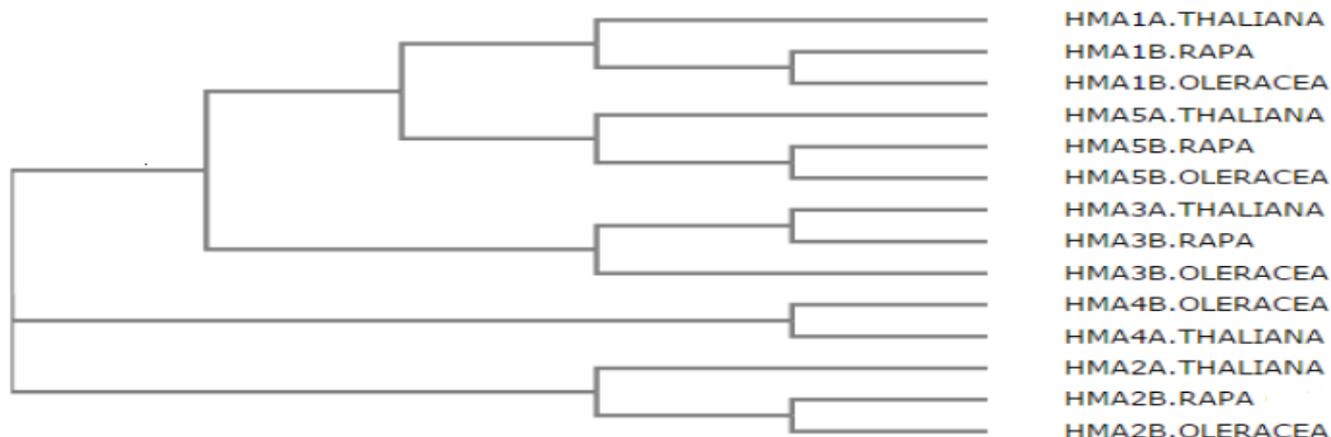
Figure2. Phylogenetic tree (cladogram) of HMA proteins with other Brassicaceae family members

3.2. Protein localization

By PSI, the proteins are localized to 10 possible locations, with a score from 0 to 1 and higher implying higher confidence in the presence of the protein in a particular subcellular compartment. Results are shown for protein localization prediction in table below:

Table3. Results of protein localization by PSI

Protein name	Predicted Subcellular Localization	Score
HMA 1	Plastid	0.689
HMA 2	Membrane	0.676
HMA 3	Membrane	0.694
HMA 4	Membrane	0.513
HMA 5	Membrane	0.393

3.3. Predicted and varified 3D structure models

The determination of the structure of proteins is vital for total understanding of the function, interactions and possible ligands, conserved domains and their homologues and many other purposes. However, experimental determination of the 3D structure is a demanding and time consuming process, so bioinformatics tools are used to predict the structures of proteins of interest. The 3D structures of HMA proteins predicted by Phyre2 tool are seen in figure 3.

Figure 3: Predicted 3D structures and respective Ramachandran plots for HMA1 protein

Figure 4: Predicted 3D structures and respective Ramachandran plots for HMA2-5 proteins.

After the prediction, the 3D structures underwent the process of validation by several structure assessment tools. The results are shown in table 4.

Table4. Structure assessment of predicted 3D structures

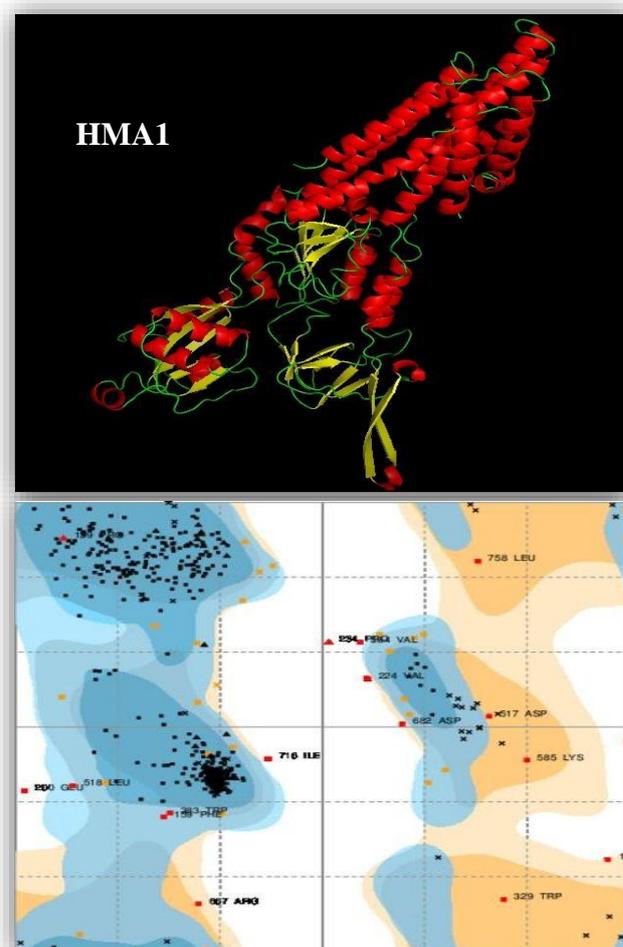| Name	QMEAN6 score	QMEAN Z-score	Verify3D	PROCHECK most favored region in %
HMA 1	0.546	-2.40	0.70	93.5
HMA 2	0.516	-2.73	0.68	90.4
HMA 3	0.523	-.2.66	0.78	92.3
HMA 4	0.554	-2.31	0.72	90.3
HMA 5	0.440	-3.57	0.80	91.1

3.4. Domain identification

The domains in the five HMA proteins from Brassica oleracea were identified by the SMART software. The results are presented in the table 5:

Table5. The domains and their positions in the five HMA proteins

Protein Names	DOMAIN NAME			
	E1-E2 ATPase		HAD	
	Start	End	Start	End
HMA 1	202	438	446	687
HMA 2	164	383	391	603
HMA 3	167	386	394	606
HMA 4	175	394	403	614
HMA 5	414	654	662	879

The SMART analysis revealed that all HMA proteins have the same domains but on different locations. The E1-E2 ATPase domain is a trans-membrane domain, which is basically membrane-bound enzyme complex/ion transporter that uses ATP hydrolysis to drive the transport of protons across a membrane. Some trans-membrane ATPases also work in reverse, harnessing the energy from a proton gradient, using the flux of ions across the membrane via the ATPase proton channel to drive the synthesis of ATP [35]. There are many different classes of P-ATPases, which transport specific types of ion: H^+, Na^+, K^+, Mg^{2+}, Ca^{2+}, Ag^+ and Ag^{2+}, Zn^{2+}, Co^{2+}, Pb^{2+}, Ni^{2+}, Cd^{2+}, Cu^+ and Cu^{2+}. P-ATPases can be composed of one or two polypeptides, and can usually assume two main conformations called E1 and E2 [36]. The HAD domain, haloacid dehydrogenase (HAD) superfamily domain, includes phosphatases, phosphonatases, P-type ATPases, beta-phosphoglucomutases, phosphomannomutases, and dehalogenases, which are involved in a variety of cellular processes ranging from amino acid biosynthesis to detoxification [37].

3.5. Interactome of HMA proteins

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is used for searching interlogs of the five proteins. The STRING database consists of known and predicted protein interactions of currently 9.6 million proteins and 2031 organisms [38]. The predicted protein interaction is classified into physical or functional associations. Before entering FASTA format of sequences into STRING, we need to specify which organism to search for our sequence. STRING doesn't offer *Brassica oleracea* as a model organism so we used organism *Brassica rapa*, because it shares more than 90 % of homology with HMA proteins from *Brassica oleracea* species.

Table6. MSA of Brassica oleracea and Brassica rapa HMA protein family (1-5)

Proteins		Similarity %
BoHMA 1	BrHMA 1	97.1
BoHMA 2	BrHMA 2	95.9
BoHMA 3	BrHMA 3	97.2
BoHMA 4	BrHMA 4	90.0
BoHMA 5	BrHMA 5	97.4

Bo: Brassica oleracea
Br: Brassica rapa

The interactome analysis revealed strong interactions of HMA1, HMA2, HMA4 proteins with FP3 (farnesylated protein 3) and FP6 (farnesylated protein 6), whereas all HMA proteins show strong interactions with ATX1 (copper metallochaperone) protein and other related copper and ion binding proteins (see supplement Table 1 and Figure 7).
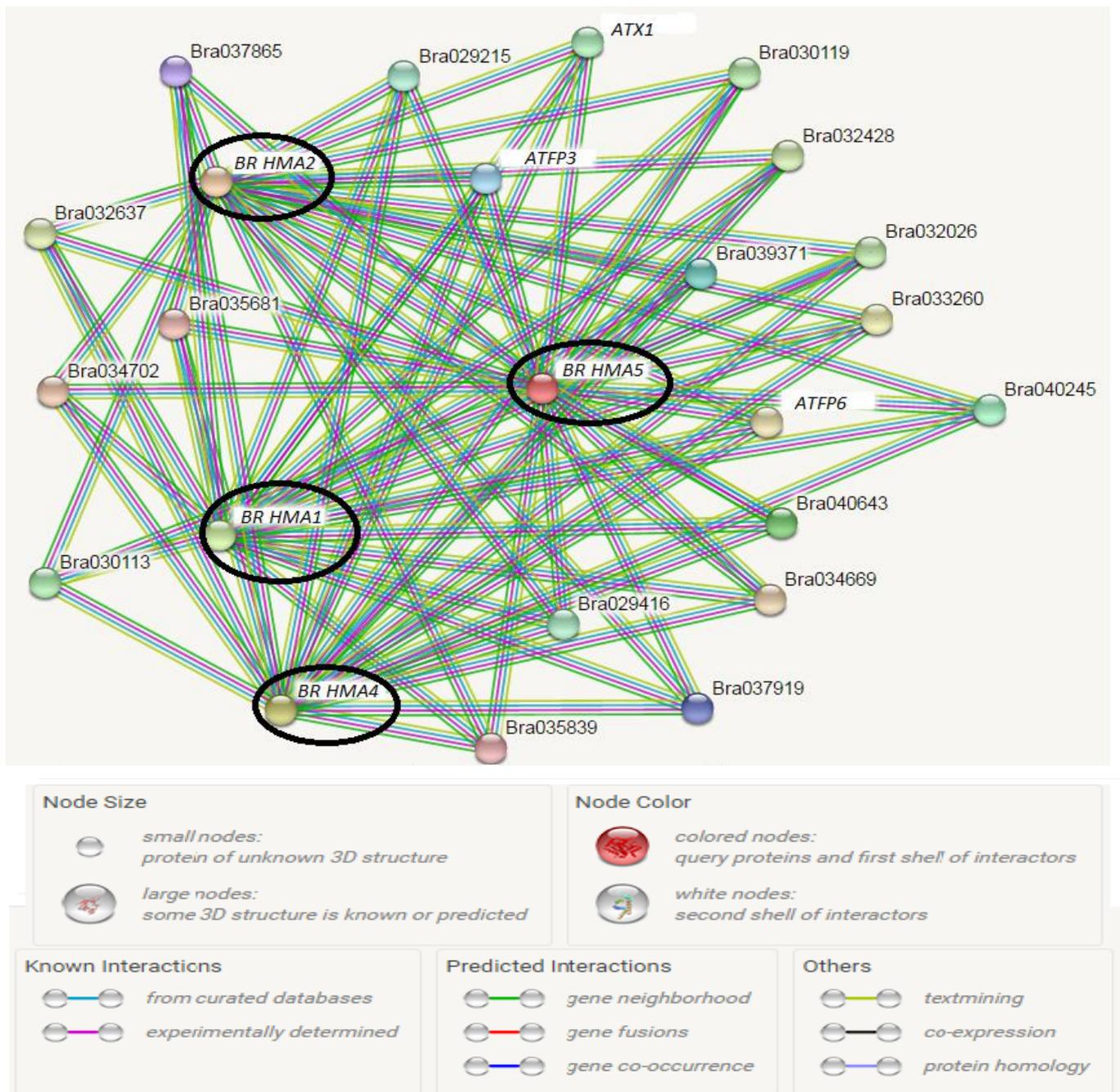
Figure 5: STRING interactome of Brassica rapa HMA1, HMA2, HMA4 and HMA5 protein family (original figure shown)

The *Brassica rapa* HMA3 protein interactome is not shown in figure 7 due to the absence of this particular protein in STRING interactome server. In order to analyze the interactome of HMA3 protein we have used the homolog from *Arabidopsis thaliana* as a template (AT4G30120), due to the high similarity of 90% with the *B.oleracea* HMA3 protein. In this analysis we have confirmed that the HMA3 from *Arabidopsis thaliana* is also predicted to interact with FP3 and FP6 as the homolog heavy metal accumulator's protein from *Brassica rapa*. Furthermore, the results show strong interactions with other metal transporting proteins such as Copper chaperone (CCH) related, heavy metal associated isoprenylated plant protein 27 (HIPP27) and several heavy metal transport/detoxification domain-containing proteins (see supplement table 2).

3.6. Docking sites prediction results

In figure 6, 7 and 8, the results of docking site prediction of HMA2 and HMA4 are shown with AtX1, FP3 and FP6, respectively.

Figure 6: Docking site prediction of ATX1 with HMA2 and HMA4 from Brassica rapa

Figure 7: Docking site prediction of FP3 and FP6 with HMA2 from Brassica rapa

Figure 8: Docking site prediction of FP3 and FP6 with HMA4 from Brassica rapa
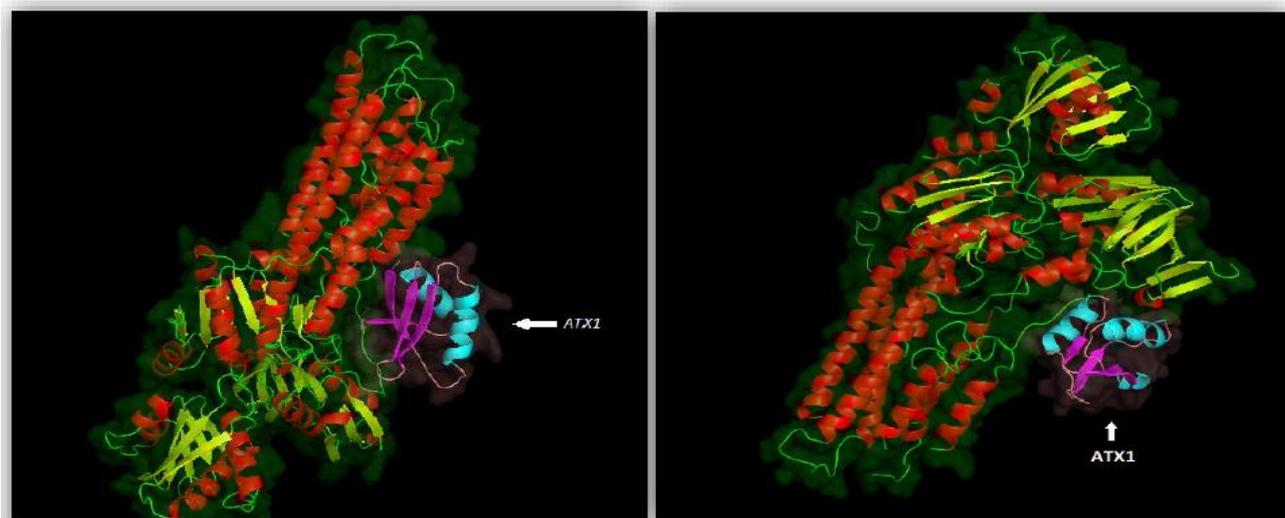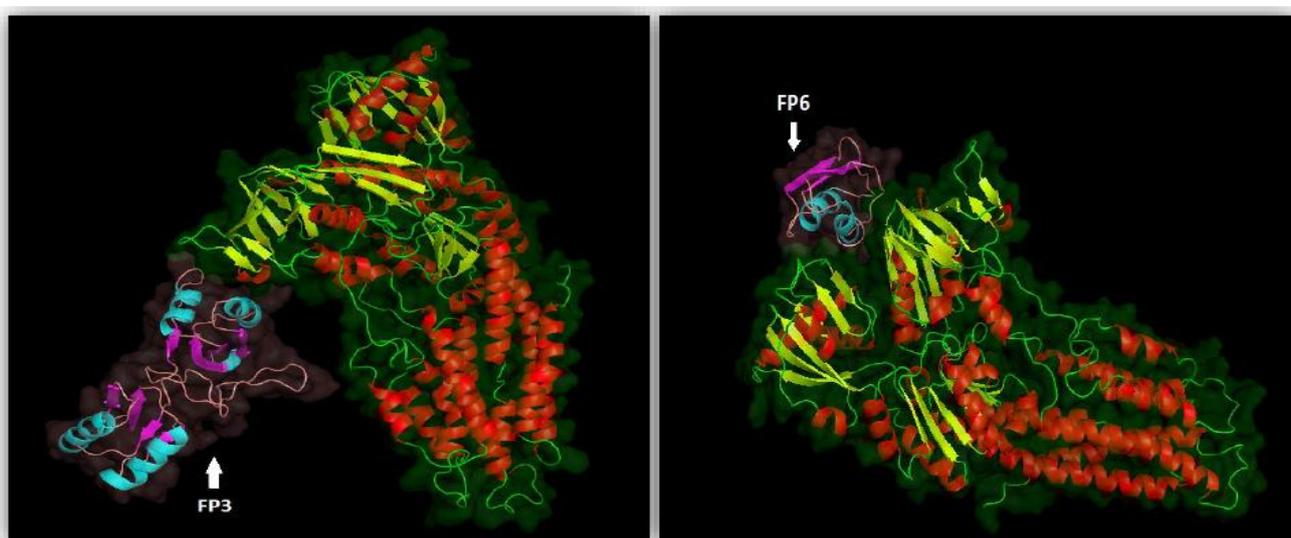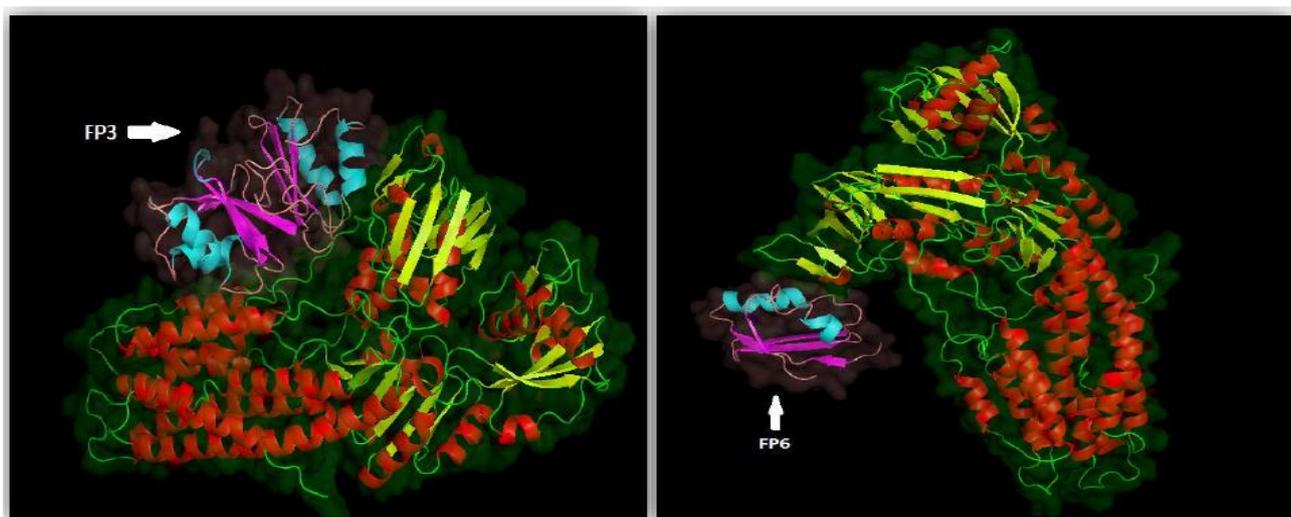
The protein structures modeled with ClusPro were checked by the same verification tools as with the HMA (1-5) proteins (Table 10). In addition, we have introduced an additional verification tool, DFire. This tool estimates the non-bonded atomic interactions in a model, thus providing the energy estimation which is considered closer to the native conformation if the DFire energy score is lower (supplement table 3) [39].

4. Discussion

Brasica oleracea is a plant known as metal hyper-accumulator that, as such can have an important role in environmental aspects. Among these, phytoremediation technology is the most interesting one and is the one that brought high attention of researchers in the last decade. However, beside phytoremediation and positive effects they can produce, metals-accumulating plants are directly or indirectly responsible for much of the dietary uptake of toxic heavy metals by humans and animals. Vegetables such as cabbage (*Brassica juncea, Brassica oleracea*) cultivated in wastewater-irrigated soils take up heavy metals in large enough quantities to cause potential health risks to the consumers [8].

Metal accumulation and translocation potential varies from plant to plant and metal to metal [8], therefore it is important to investigate both potentials in plants which are considered as metal-accumulators, which, in our work was *B.oleracea.*

Analysis of proteins responsible for metal accumulation and transport is of great importance to understand how those plants perform their functions in hyper-accumulation of metals. The *HMA* proteins (1-5) analyzed in our work are already known to play important roles in heavy metal accumulation processes. Beside their primary function as metal-accumulators, it is important to investigate other processes in which *HMA*s can be involved. In order to investigate such processes, we were looking for potential interactions with other proteins that are currently unknown to the literature, not known to interact with the analyzed *HMA* proteins.

In this work we have confirmed that *HMA2* and *HMA4* proteins share the most homology among other HMA family proteins [40], with 71.7% similarity. The phylogenetic tree analysis between the *HMA* proteins, additionally confirmed the similarity among these two proteins, where *HMA2* and *HMA4* share same ancestor, separated by other groups in the tree. These results suggest that due to their close evolutionary relationship, they play important biochemical roles by performing same or similar functions within the cell.
As reviewed by Hussain and colleagues [41], *HMA2* and *HMA4* play an important role in Zn transport and

homeostasis in *A.thaliana*. By mutating the *HMA4* and *HMA2* genes they have observed a significant decrease in Zn accumulation. Furthermore, they observed that only the *hma2-hma4* double mutant and neither of the single mutants exhibited an obvious nutritional deficiency in soil, suggesting that *HMA2* and *HMA4* have a level of functional redundancy, which can be consistent with sequence comparisons that show that *HMA4* is the most closely related to *HMA2*, as confirmed in this study.

For further phylogenetic investigation for *HMA* homologes from the Brassicaceae family, a ClustaOmega cladogram was constructed (Figure 5), revealing clear and strong phylogenetic relationship between the target species (*B.oleracea)* and other Brassicacea family species, *Arabidopsis thaliana* and *Brassica rapa.*

Due to high sequence similarity (with more than 90%) and common conserved domains of all HMA proteins of *B.oleracea* with *B.rapa* and *A.thaliana* HMA proteins, we may conclude that all know functions observed in *A.thaliana* and *B.rapa* HMA proteins can be attributed to *B.oleracea* homologues HMA proteins. As it is seen in Figure 5, *A.thaliana* HMA proteins share the common ancestors with all other taxa, indicating the possibility of an evolutional change of *A.thaliana* HMAs into *B.rapa* and *B.oleracea* HMA proteins. The HMA proteins from *B.rapa* and *B.oleracea* share a common ancestor, being sister taxa with all HMA proteins, except HMA3. HMA3 proteins from *B.rapa* and *A.thaliana* share the common ancestor, whereas *B.Oleracea* has evolved separately.

The 3D structures of proteins enables additional functional studies, domain analysis, molecular interaction studies, estimation of structural similarity between proteins etc. In this study, we used Phyre2 tool, a protein homology modeling server, used to create models of target proteins. These models contain information about the tendency for mutation of each amino acid in a sequence and are unique for each protein. They are created for a set of known 3D structures as well as for the user sequence, and then scanned to find a match [42]. Further confirmation and verification of the modeled structure was tested by three validation methods. QMEAN6, PROCHECK and Verify 3D.

The verification results of all five HMA protein in *B.oleracea,* showed sufficient quality, required for further analysis. According to our results, the Verify3D score for HMA1 is 0.70, for HMA2 it is 0.68, for HMA3 0.78, for HMA4 it is 0.72 and the highest score was observed with HMA5, being 0.80. The Ramachandran plots analysis revealed that all of the structural regions lie in the range of acceptance, with having more than 90% favored regions. QMEAN6

results revealed good Z scores. To be precise, QMEAN6 score for HMA1 is 0.546, for HMA2 score is 0.516. 0.523 is score for HMA3 while 0.554 is score for HMA4. The lowest score is 0.444 and it stand for HMA5 (see Table 5). For the Z-score analysis we observe that all models have negative Z-scores (in average of -2 Z-score), being a median score for structural validations. Models of low quality are expected to have strongly negative QMEAN Z-scores, less then -3.5 [43]. Obtained Z-scores are in line to scores obtained for high-resolution experimental structures of similar sizes solved by X-raycrystallography. Therefore, the Phyre2 generated models appeared acceptable for the protein and metal docking site prediction.

We have shown that all proteins share two domains, starting from different residues. The domains identified are P-ATPases (E1-E2 ATPases), membrane-bound enzyme complexes/ion transporters that use ATP hydrolysis to drive the transport of protons across a membrane [44] and the HAD domain, haloacid dehydrogenase (HAD) superfamily domains which are involved in a variety of cellular processes ranging from amino acid biosynthesis to detoxification [37]. The interactome analysis revealed strong interactions of HMA1, HMA2, HMA4 proteins with FP3 (farnesylated protein 3) and FP6 (farnesylated protein 6), whereas all HMA proteins show strong interactions with ATX1 (copper metallochaperone) protein and other related copper and ion binding proteins. In order to analyze the interactome of HMA3 protein we have used the homolog from *A.thaliana* as a template (AT4G30120), due to the high similarity of 90% with the *B.oleracea* HMA3 protein. The HMA3 protein from *A.thaliana* is also predicted to interact with FP3 and FP6 as the homolog heavy metal accumulator's protein from *B.rapa*. Furthermore, the results show strong interactions with other metal transporting proteins such as Copper chaperon (CCH) related, heavy metal associated isoprenylated plant protein 27 (HIPP27) and several heavy metal transport/detoxification domain-containing proteins (supplement table 1).

FP6 (also known as HIPP26) is characterized by a heavy metal binding domain (HMA) and an additional isoprenylation motif on C-terminus. This family of HIPPs embraces at least 44 proteins in *A.thaliana* with HMA domain being responsible for heavy metal binding, metal transport and metal homeostasis processes. Isoprenylation motif is added through the process of isporenylation [45].

Isoprenylation, also known as farnesylation, is a post-translational protein modification that involves addition of a C–terminal hydrophobic anchor that is important for interaction of the protein with membranes or other proteins [46].

In a study conducted by Barth and colleagues [45], it is confirmed that HIPP26 exhibits a nuclear localization signal (NLS), thus being localized in the nucleus. In their work, they also concluded that for the exact spatial localization of HIPP26 within the nucleus, the isoprenylation seems to be important, which probably by its hydrophobic nature determines the correct spatial arrangement of this protein within the nucleus.

Furthermore, their study confirmed that HIPP26 strongly interacts with ATHB29, a zinc fingerhomeodomain transcription factor (ZF-HD proteins) which is found to be induced by drought, high salinity and abscisic acid, thus playing role in regulation of stress response of plants [45]. Furthermore, Gao and colleagues [47] showed that FP6 in A.thaliana (AtFP6) upon interaction with plasma mebrane acyl-CoA-binding protein 2 (ACBP2) mediate cadmium Cd(II) tolerance [47].

Due to the strong interaction with FP6, the represented data confirms HMA protein family involvement in Cd(II) transport and tolerance, since all three HMAs are found to be cadmium/zinc transporting ATPases. In addition, we may suggest that these three HMA proteins may be important in strees-induced tolerance, since it was the case for FP6 protein [45]. FP3 from *A.thaliana*, if soluble and isoprenylated, is capable of reversibly binding a copper-chelate matrix in tobacco BY2 cell homogenates, suggesting a ubiquitous role for these proteins in diverse plants [48]

In this study, we confirm the interaction of all *HMA* proteins with CCH (Copper chaperone) or CCH-related proteins, which has been shown to functionally complement atx1 mutants, but the ATFP3 gene expression is not regulated in the same manner as CCH gene expression [49].

ATX1 (copper metallochaperone) protein shows strong interactions with *HMA1*, *HMA2, HMA4* and *HMA5* proteins. ATX1 is related with copper metallochaperones which assist copper in reaching vital destinations without inflicting damage or becoming trapped in adventitious binding sites [50]. ATX1 is shown to bind Cu(I) in the cytoplasm which delivers it to a copper transporter in the membrane of a post-Golgi vesicles. In the vesicle, the copper is inserted into a multicopper oxidase essential for high-affinity iron uptake, so ATX1 can be involved in both, copper transport and defense against oxidative stress [49] ATX1 is also proposed to be involved in Cu homeostasis by its Cu-binding activity and interaction with the Cu transporter heavy metal-transporting P-type ATPase5, suggesting a regulatory role for the plant-specific domain of the CCH Cu chaperone, therefore, a role for *HMA5* in Cu compartmentalization and detoxification [21].

In a more recent study conducted by Lung et al. [51], it is confirmed that overexpression of ATX1 enhancing Cu tolerance implies the potential use of ATX1 for phytoremediation in Cu-contaminated soil. In same study, they connected *HMA5* with ATX1 on the way that ATX1 was proposed to deliver Cu to HMA5 for Cu detoxification in roots and translocation to shoots.

In the docking analysis, the HMA proteins were considered to be a ligand (according to ClusPro default settings, the ligand is the structure that gets rotated to fit into the receptor). Specific docking sites presented are ATX1 with *HMA2* and *HMA4*. In this research we have verified our structures where all predicted models of docking show sufficient quality (see supplement table 3).

For further analyses of predicted docking structures, the electrostatic potential between the *HMA* proteins and docking partners was calculated via DeepView. This tool is showing clouds of negative and positive electrostatic potential in the docking site predicted, from which we could conclude that at least part of all docking sites is due to electrostatic forces. In supplement figure 1 we can see the clear separation of charges between *HMA2 /HMA4* and ATX1 on similar docking regions.

The predicted docking region lies in the of N terminun as shown in literature, where *HMA2* and *HMA4* N-terminal domain are essential for function *in planta* while the C-terminal domain, although not essential for function, may contain a signal important for the subcellular localization of the protein (supplement figure 2)[52].

These predicted docking sites of FP3 and FP6 to *HMA2* and *HMA4* lie in similar region, usually on N terminus, which confirms the good modelling of 3D structures by ClusPro. The visualization of the resulting docking site models for *HMA4* with FP3 and PP6 (supplement figure 3).

All the structures, verified by the electrostatic potential, given by DeepView, confirm the docking sites supported by the electrostatic forces. It has been shown that the electrostatic potentials at the interfaces of interacting molecules are anti-correlated. This means that at the interface, there is a good chance to find a patch of positive electrostatic potential on the surface of one molecule positioned next to a negative patch on the surface of the adjacent molecule and vice versa [53]. Furthermore, a big DFire score results for all models are indicating good models of docking, which estimates the non-bonded atomic interactions in a model, thus providing the energy estimation that is closer to the native conformation the lower it gets (lower quality predictions have more negative scores) [39].

5. Conclusion

Brassicaceae family plants are known to accumulate high amounts of toxic metals, such as: (N), phosphorus (P), potassium (K), calcium (Ca), magnesium (Mg) and sulphur (S) and the micronutrients boron (B), chlorine (Cl), iron (Fe), manganese (Mn),copper (Cu), zinc (Zn), nickel (Ni) and molybdenum (Mo). The subgroup, *Brassica oleracea* shows great potential to be hyper-accumulator of Zn and Cd heavy metals. In this research we have focused to further investigate the roles of *HMA* (1-5) proteins in *Brassica oleracea*, predicting their structures and interactomes. To achieve the most accurate result, the stated aim was enhanced and supported through the use of several common bioinformatics techniques. The *B.oleracea HMA* proteins (1-5) were subjected to multiple sequence alignment analysis with *HMA* proteins from *B.rapa, B.napus and A.thaliana* in order to obtain information about conserved regions among these proteins and to assess the phylogenetic relationship of the proteins. This further enabled further analysis of their 3D structures as well as their interactome analysis, in order to confirm current functional roles of each protein and possibly discover new inteactome partners unknown to the literature, for new annotations of functional roles of *B.oleracea HMA* proteins (1-5).

It is through bioinformatics analysis that we identified and structurally predicted 5 homologues of *HMA* proteins in *Brassica rapa,* mostly similar to *Brassica oleracea*. Since they are similar, but not identical in structure and differentiate in two groups in phylogenetic analysis, further inference about the functions and localization of the homologues was required. For that purpose, localization tools were used to predict subcellular locations and the trend of differences between the homologues continued. Lastly, the interactome analysis showed similar functions and associations with many crucial processes of metal ion transportation, required for cellular integrity and stability maintenance. The results obtained in this study lead us to the conclusion that cellular functions of the 5 homologues are very similar, where *HMA2* and *HMA4* are directly involved in Zn/Cd transport, whereas *HMA5* functions as metallochaperones and functions in copper detoxification, as confirmed within this study. Furthermore, *HMA2* and *HMA4* are shown to have strong interaction with ATX1 protein, by now only know to interact with *HMA5*, which may indicate a specific involvement of *HMA2* and *HMA4* proteins in Cu(I) binding and the delivery to the post- Golgi vesicle, with strong possibility of Cu compartmentalization and detoxification, as shown for *HMA5*.

In addition, the interactome analysis revealed strong interaction of *HMA2* and *HMA4* proteins with FP3 (farnesylated protein 3) and FP6 (farnesylated protein 6). A study conducted by Dykema et al. [48], showed that FP3 from *A.thaliana* has function as ubiquitous protein in diverse plants. FP6 (also known as HIPP26) is characterized by a heavy metal binding domain (HMA) and an additional isoprenylation motif on C-terminus. It is shown that HIPP26 strongly interacts with ATHB29, a zinc finger homeodomain transcription factor (ZF-HD proteins) which is found to be induced by drought, high salinity and abscisic acid, thus playing role in regulation of stress response of plants. Furthermore it is shown that FP6 interacts with acyl-CoA-binding protein 2 (ACBP2) mediate cadmium Cd(II) tolerance protein, indicating the possibility the *HMA2* and *HMA4* proteins may share the above mentioned cellular functions.

Experimental determination of 3D structures of the homologues, as well as further testing in terms of interactome and co-localization analysis, is needed to fully understand the role of *HMA* homologues in metal transports. Especially the docking sites and binding domains need to be researched further, preferably *in vivo*, in order to understand the mechanism by which this protein docks to Zn and Cd ions and its function as a partner for other protein functions. This study confirmed the known functional roles of *HMA* proteins, especially the *HMA2* and *HMA4* proteins, known to be hyper-accumulators for Zn and Cd, elongating their potential cellular roles by detailed 3D structure and interactome analysis.

References

[1] A.M. Gardner, A. F. Brown, J. A. Juvik, "QTL analysis for the identification of candidate genes controlling phenolic compound accumulation in broccoli (Brassica oleracea L. var. italica)," *Mol Breeding,* vol. 36:81, 2014.

[2] M. P. Mourato, I. N. Moreira, L. Leitão et al., "Effect of Heavy Metals in Plants of the Genus Brassica," *International Journal of Molecular Sciences,* ISSN 1422-0067, pp 1-24, 2015.

[3] H. J. Korry, J. W. Finley, A. Sigrid-Keck, R. J. Robbins," *The Journal of Nutrition,* vol. 135, no. 5, 2005.

[4] L. Fengjuan, W.Yongsheng, F. Xiaoling, J. Yucui, W. Zhaojia, H. William, P. Carolyn, L. Lei et al., "A Novel Mechanism of Indole-3-Carbinol Effects on Breast Carcinogenesis Involves Induction of Cdc25A Degradation," *Cancer Prevention Research,* vol. 3, no. 7, 2010.

[5] M. W. Farnham and D.A. Kopsell, "Importance of Genotype on Carotenoid and Chlorophyll Levels in Broccoli Heads," *HortScience,* vol. 44, no. 5, 2009.

[6] B. Srilakshmi, *Nutrition Science*. New Age International. ISBN 978-81-224-1633-6, pp. 186–7, 2006.

[7] S. Christensen, R. von Bothmer, P.Gert et al., "AFLP analyses of genetic diversity in leafy kale (Brassica oleracea L. convar. acephala (DC.) alef) landraces, cultivars ad wild population in Europe,'' *Genet Reseour Crop Evol,* vol. 58, 2010.

[8] G.A. Boamponsem, M. Kumi and I. Debrah, "Heavy Metals Accumulation In Cabbage, Lettuce And Carrot Irrigated With Wastewater From Nagodi Mining Site In Ghana," *International Journal of Scientific and Technology Research,* vol. 1, no. 11, pp. 124-129, 2012.

[9] S. Mohammad, A. Khan, Z.R. Goel and J. Musarrat, "Biomanagement of Metal-Contaminated Soils," In: Environmental Pollution. Springer Netherlands (online), 2011.

[10] L. Subasic, H. Gavranovic, I. Muhovic and A. Memon, "Heavy metal induced gene expression in Brassicaceae," Department of Genetics and Bioengineering, Faculty of Engineering and Information Technologies, International Burch University, BiH, 2012.

[11] M. Cho, C.N. Agnes and D. Karl-Josef, "Differential heavy metal tolerance of Arabidopsis halleri and Arabidopsis thaliana: a leaf slice test," *The New Phytologist,* vol. 158, no. 2, pp. 287-293, 2002.

[12] V. Shanmugam, Lo, Jing-Chi and Kuo-Chen. Yeh, "Control of Zn uptake in Arabidopsis halleri: a balance between Zn and Fe," *Frontiers in Plant science,* vol. 4, pp. 281, 2013.

[13] X. Sun, G.Yu, Jing-Tao. Li et al., "A Heavy Metal-Associated Protein (AcHMA1) from the Halophyte, Atriplex canescens (Pursh) Nutt., Confers Tolerance to Iron and Other Abiotic Stresses When Expressed in Saccharomyces cerevisiae," *Int J Mol Sci.,* vol. 15, no. 8, pp. 14891–14906, 2014.

[14] J. M. Argüello, E. Eren, M.González-Guerrero, "The structure and function of heavy metal transport P1B-ATPases," *M. Biometals,* vol. 20, 2007.

[15] E. Eren and José M. Argüello, "Arabidopsis HMA2, a Divalent Heavy Metal-Transporting PIB-Type ATPase, Is Involved in Cytoplasmic Zn2+ Homeostasis," *Plant Physiol.,* vol. 136, no. 3, pp. 3712–3723, 2004.

[16] P. Ashton and K. Leon, "Identification of Thlaspi caerulescens Genes That May Be Involved in Heavy Metal Hyperaccumulation and Tolerance. Characterization of a Novel Heavy Metal Transporting ATPase, " *Plant Physiology*, vol. 136, no. 3, pp. 3814-3823, 2004.

[17] K. Ute, I. N. Talke and M. Hanikenne, "Transition metal transport," *FEBS Letters,* vol. 581, no. 12, pp. 2263–2272, 2007.

[18] O. Siemianowski, A. Barabasz, M. Kendziorek, E. Bulska, L. E Williams and D.M. Antosiewicz, "HMA4 expression in tobacco reduces Cd accumulation due to the induction of the apoplastic barrier," J *Exp Bot.,* vol. 65, no. 4, 2014.

[19] A. Gravot, A. Lieutaud, F. Verret, P. Auroy, A. Vavasseur and P. Richaud, "AtHMA3, a plant P1B-ATPase, functions as a Cd/Pb transporter in yeast," *FEBS Lett.,* vol. 561, 2004.

[20] Y. Kobayashi, K. Kuroda, K. Kimura et al., "Amino acid polymorphisms in strictly conserved domains of a P-type ATPase HMA5 are involved in the mechanism of copper tolerance variation in Arabidopsis," *Plant Physiol.,* vol. 148, no. 2, 2008.

[21] A. S. Colás, V. S. Rodríguez-Navarro et al., "The Arabidopsis heavy metal P-type ATPase HMA5 interacts with metallochaperones and functions in copper detoxification of roots," *The Plant Journal,* vol. 45, no. 2, pp. 225–236, 2005.

[22] N. Grotz and M. L. Guerinot, "Molecular aspects of Cu, Fe and Zn homeostasis in plants," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research,* vol. 1763, no. 7, pp. 595–608, 2006.

[23] E.W. Sayers, T. Barrett, S.H. Benson, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res.,* vol. 37, D5-15, 2009.

[24] M. Goujon, H. McWilliam, W. Li et al., "A new bioinformatics analysis tools framework at EMBL-EBI," *Nucleic acids research,* vol. 38. Suppl: W695-9, 2010.

[25] A. M. Lesk, Introduction to Bioinformatics. Oxford University Press Inc., New York, USA, 2002.

[26] M.A. Larkin, B. G. Blackshields, N.P. Brown et al., "ClustalW and ClustalX version 2," *Bioinformatics,* vol. 23, no. 21, pp. 2947-2948, 2007.

[27] L.A. Kelley and M.J.E. Sternberg, "Protein structure prediction on the web: a case study using the Phyre server," *Nature Protocols,* vol. 4, pp. 363 – 371, 2009.

[28] P. Benkert, M. Biasini and T. Schwede, "Toward the estimation of the absolute quality of individual protein structure models," *Bioinformatics,* vol. 27, no. 3, 2011.

[29] D. Eisenberg, R. Lüthy and J.U. Bowie, "Assessment of protein models with three-dimensional profiles," *Nature,* vol. 356, no. 6364, 1992.

[30] R.A. Laskowski, M.W. MacArthur, D.S. Moss and J.M. Thornton, "PROCHECK - a program to check the stereochemical quality of protein structures," *J. App. Cryst.,* vol. 2, pp. 283-291, 1993.

[31] L. Liu, Z. Zhang, Q. Mei and M. Chen, "PSI: A Comprehensive and Integrative Approach for Accurate Plant Subcellular Localization Prediction," *PLoS ONE,* vol. 8, no.10, 2013.

[32] J. Schultz, F. Milpetz, P. Bork and C.P. Ponting, "SMART, a simple modular architecture research tool: Identification of signaling domains," *Proc. Natl. Acad. Sci.*, vol. 95, no. 11, pp. 5857-5864, 1998.

[33] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: proteinprotein interaction networks, integrated over the tree of life," *Nucleic Acids Res.,* vol. 43, 2015.

[34] D. Kozakov, D. Beglov, T. Bohnuud et al., "How good is automated protein docking?" *Proteins,* vol. 81, no. 12, pp. 2159-2166, 2013.

[35] R.L. Cross and V. Muller, "The evolution of A-, F-, and V-type ATP synthases and ATPases: reversals in function and changes in the H+/ATP coupling ratio," *FEBS Lett*., vol. 576, pp. 1-4, 2004.

[36] K. B. Axelsen and M.G. Palmgren, "Evolution of substrate specificities in the Ptype ATPase superfamily," *J Mol Evol.,* vol. 46, no. 1, pp. 84-101, 1998.

[37] E.V. Koonin and R.L. Tatusov, "Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity.

Application of an iterative approach to database search," *J Mol Biol,* vol. 244, no. 1, 1994.

[38] I. Letunic, T. Doerks and P. Bork, "SMART: recent updates, new developments and status in 2015," *Nucleic Acids Res.,* vol. 43, 2015.

[39] H. Zhou and Y. Zhou, "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction," *Protein Sci.,* vol. 11, pp. 2714-2726, 2002.

[40] K. Viehweger, "How plants cope with heavy metals," *Botanical Studies,* vol. 55, no. 35, 2014.

[41] H. D. Haydon, M. J. Wang et al., "P-Type ATPase Heavy Metal Transporters with Roles in Essential Zinc Homeostasis in Arabidopsis," *Plant Cell,* vol. 16, no. 5, pp. 1327–1339, 2004.

[42] L. A. Kelley et al., "Phyre2 Protein Fold Recognition Server," *Nature Protocols,* vol. 10, pp. 845-858, 2015.

[43] P. Benkert, M. Biasini and T. Schwede, "Toward the estimation of the absolute quality of individual protein structure models," *Bioinformatics,* vol. 27, no. 3, 2011.

[44] K. B. Axelsen and M.G. Palmgren, "Evolution of substrate specificities in the Ptype ATPase superfamily," *J Mol Evol.,* vol. 46, no. 1, pp. 84-101, 1998.

[45] B. Olaf, S. Vogt, R. Uhlemann, Z. Wiebke and H. Klaus, "Stress induced and nuclear localized HIPP26 from Arabidopsis thaliana interacts via its heavy metal associated domain with the drought stress related zinc finger transcription factor ATHB29," *Plant Mol Biology,* vol. 69, pp. 213–226, 2009.

[46] Braga de Abreu-Neto, A. C. Joao, L.F.V. Turchetto-Zolet et al., "Heavy metal-associated isoprenylated plant protein (HIPP): characterization of a family of proteins exclusive to plants," *The FEBS Journal,* vol. 280, no. 7, 2013.

[47] G, Wei, Y.L. Hong, X. Shi and Mee-Len Chye, "Protein interactors of acyl-CoA-binding protein ACBP2 mediate cadmium tolerance in Arabidopsis," *Plant Signal Behav.,* vol. 5, no. 8, pp. 1025–1027, 2010.

[48] E. Dykema, R. Sipes, A. Marie et al., "A new class of proteins capable of binding transition metals," *Plant Mol Biol.,* vol. 41, no. 1, 1999.

[49] E. Himelblau, H. Mira, S. Lin et al., "Identification of a Functional Homolog of the Yeast Copper Homeostasis Gene ATX1 from Arabidopsis," *Plant Physiol.,* vol. 117, no. 4, pp. 1227–1234, 1998.

[50] N. J. Robinson and Dennis R. Winge, "Copper Metallochaperones," *Annu Rev Biochem.,* vol. 79, pp. 537–562, 2010.

[51] S. Lung-Jiun, L. Jing-Chi and Y. Kuo-Chen, "Copper Chaperone Antioxidant Protein1 Is Essential for Copper Homeostasis," *Plant Physiol.,* vol. 159, no. 3, pp. 1099–1110, 2012.

[52] K.C. Wong, E. Jarvis, S. Renée et al., "Functional analysis of the heavy metal binding domains of the Zn/Cd-transporting ATPase, HMA2, in Arabidopsis thaliana," *New Phytologist Journal,* vol. 181, no. 1, pp. 79–88, 2009

[53] A. Heifetz, E. Katchalski-Katzir and M. Eisenstein, "Electrostatics in protein–protein docking," *Protein Sci.,* vol. 11, no. 3, pp. 571–587, 2002.
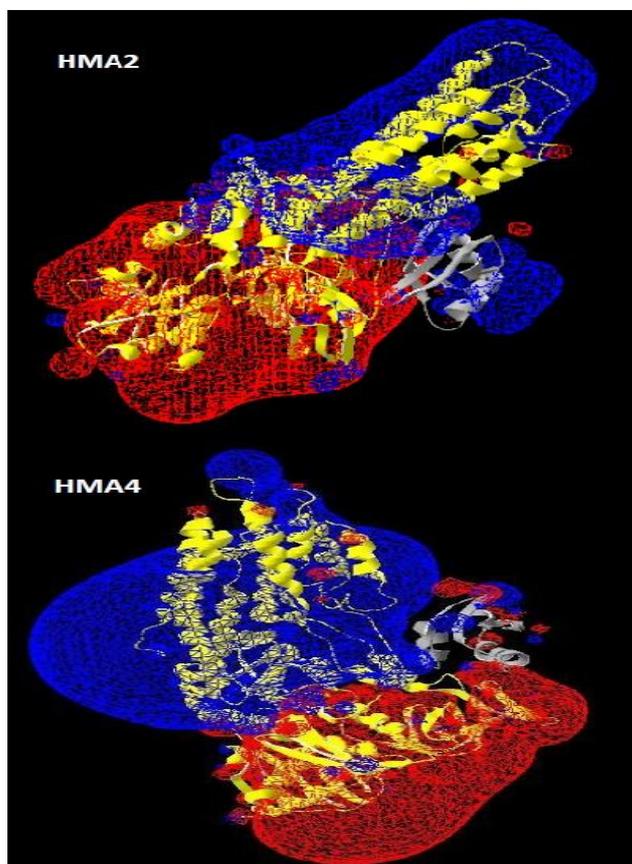
Supplement Figure 1: Visualization of the electrostatic potential of the *HMA2* and *HMA4* with ATX1 docking site as modeled by ClusPro. Yellow=HMA ribbon, grey= ATX1 ribbon, red=negative potential, blue=positive potential
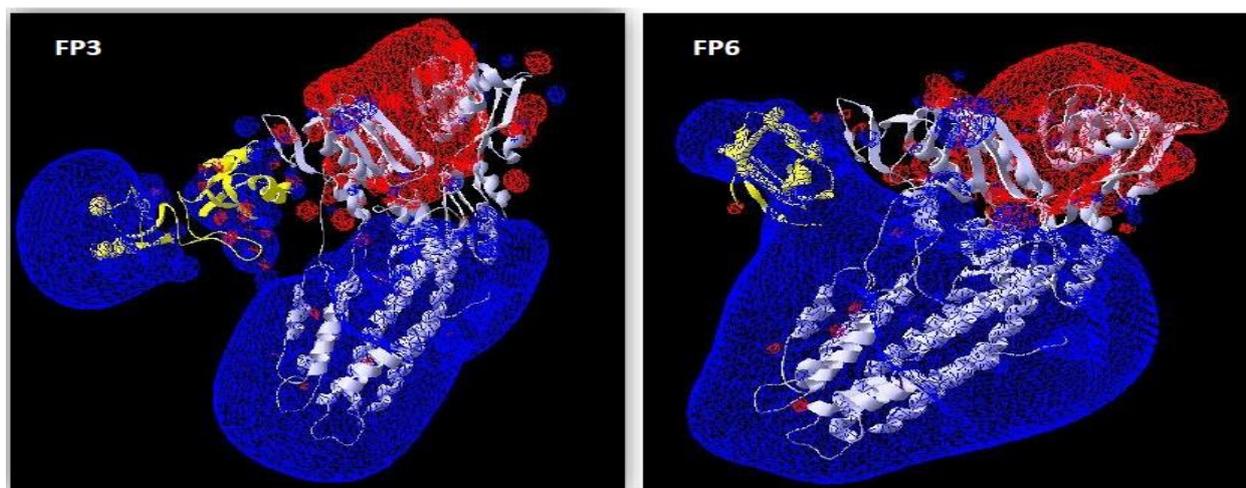
Supplement Figure 2: Visualization of the electrostatic potential of the FP3 and FP6 with *HMA2* docking site as modeled by ClusPro. Yellow= FP3/6 ribbon, grey= HMA2 ribbon, red=negative potential, blue=positive potential
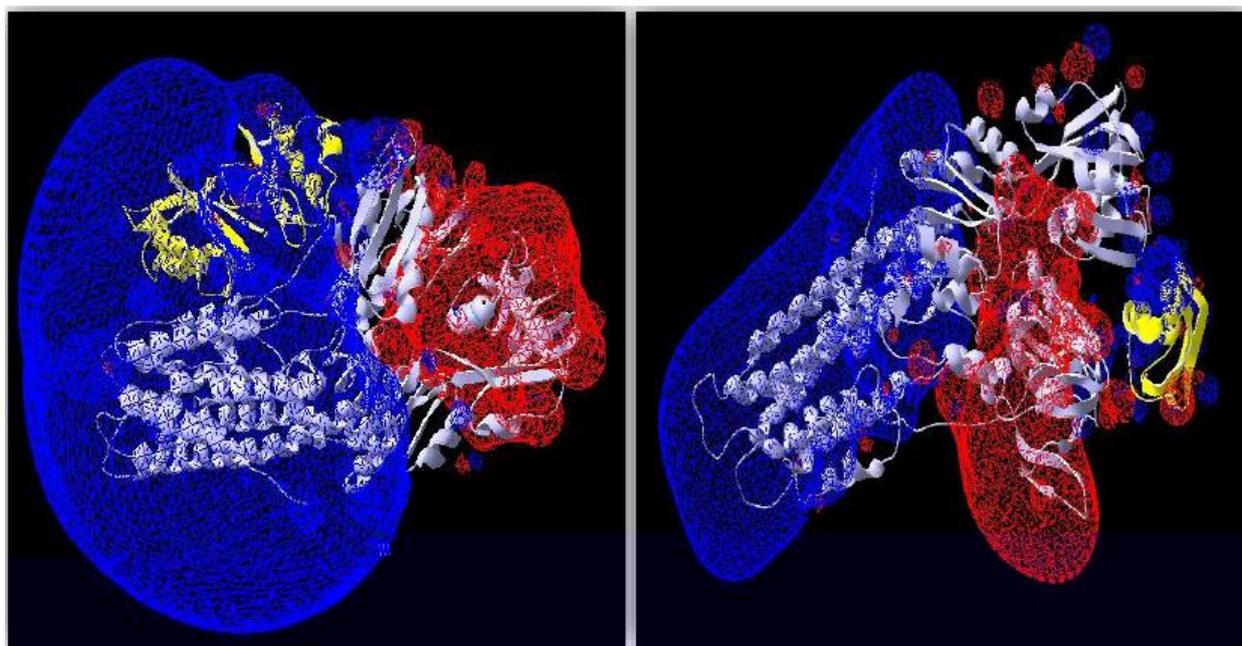
Supplement Figure 3: Visualization of the
electrostatic potential of the FP3 and FP6 with *HMA4* docking site as modelled by ClusPro. Yellow= FP3/6 ribbon, grey= HMA2 ribbon, red=negative potential, blue=positive potential

Supplement table 1: Interactome results of Brassica rapa HMA1, HMA2, HMA4 and HMA5 protein family

HMAs	Scores	Brassica rapa interactome ID	Brassica rapa STRING Annotation	Function(s)
HMA1	0.974	Bra035681	Copper-binding family protein	Copper ion binding
HMA4				
HMA1	0.974	Bra033260	Heavy-metal-associated domain-containing protein	Metal ion binding
HMA2				
HMA4				
HMA1	0.975	Bra029854	Arabidopsis homolog of anti-oxidant 1-ATX1	Chaperon protein. Predominant role in delivery of Cu to HMA and Cupper homeostasis.
HMA2				
HMA4				
HMA5				
HMA1	0.974	Bra032026	Copper-binding protein-related	Copper ion binding
HMA2				
HMA4				
HMA1	0.975	Bra039371	Hydroxyproline-rich glycoprotein family protein	Involved in copper import and transfer through cells
HMA2				
HMA4				
HMA5				
HMA1	0.975	Bra037865	T15F16.6	Heavy-metal-associated domain-containing protein / copper chaperone (CCH)-related
HMA2				
HMA4				
HMA5				
HMA1	0.974	Bra037919	Metal ion binding	Metal ion binding

HMA	Scores	Interactome TAIR ID	STRING Annotation	Functions
HMA2				
HMA4				
HMA5				
HMA1	0.974	Bra033570	HIPP26/FP6/ FARNESYLATED PROTEIN 6	Heavy-metal-binding protein, heat acclimation, Binds lead, cadmium and copper.
HMA2				
HMA4				
HMA1	0.974	Bra038642	FP3/ FARNESYLATED PROTEIN 3	Heavy-metal-binding protein. Binds lead, cadmium and copper. May be involved in heavy-metal transport
HMA2				
HMA4				

Supplement table 2: *Interactome results of Arabidopsis thaliana HMA3 protein*

HMA	Scores	Interactome TAIR ID	STRING Annotation	Functions
ATHMA3	0.961	AT3G56240.1	Copper chaperon (CCH) related.	Stress inducing. Metal ion binding
	0.961	AT5G02600.1	Sodium Possium Root Defective -NAKKR1	Heavy metal transport/detoxification superfamily protein1
	0.961	AT5G66110.1	HIPP27	Heavy metal associated isoprenylated plant protein 27.Metal Ion Binding
	0.961	AT5G63530.1	Farnesylated protein 3- FP3/ MLE2_16	Heavy-metal-binding protein. Binds lead, cadmium and copper. May be involved in heavy-metal transport
	0.961	AT4G38580.1	Farnesylated protein 6 - FP6/HIPP26	Heavy-metal-binding protein, heat acclimation, Binds lead, cadmium and copper.
	0.961	AT5G60800.1	Heavy metal transport/detoxification domain-containing proteins	Heavy metal transport/detoxification domain-containing protein
	0.961	AT5G37860.1		
	0.961	AT5G27690.1		
	0.961	AT4G39700.1		
	0.961	AT5G03380.1		
	0.961	AT5G19090.1		
	0.961	AT5G24580.1		

Supplement table 3: Verification analysis of all chosen docking candidates with HMA2 and HMA4

Name	QMEAN6 score	QMEAN6 score Z-score	DFire energy	PROCHECK most favored region in %
HMA2-ATX1	0.511	-2.795	-1007.50	85.0
HMA2-FP6	0.454	-3.413	-1006.70	84.2
HMA2-FP3	0.313	-4.979	-1124.00	82.08
HMA4-FP3	0.537	-4.484	-1111.77	83.0
HMA4-FP6	0.454	-3.413	-1006.70	84.2
HMA4-ATX1	0.484	-3.091	-996.22	84.9

\