SpADS and SNAP-NAPPA Microarrays towards Biomarkers Identification in Humans: Background Subtraction in Mass Spectrometry with E.coli Cell Free Expression System

Claudio Nicolini1,*, Rosanna Spera1 and Eugenia Pechkova2
1Nanoworld Institute Fondazione EL.B.A. Nicolini, Largo Piero Redaelli 7, Pradalunga (Bg), Italy
2Laboratories of Biophysics and Nanobiotechnology, Department Experimental Medicine, University of Genova, Italy

*Corresponding author: Claudio Nicolini, President Nanoworld Institute Fondazione EL.B.A Nicolini, Italy, 1-650-268-9744; Fax: 1-650-618-1414; E-mail: president@fondazioneelba-nicolini.org

Rec date: Sep 29, 2014; Acc date: Jan 15, 2015; Pub date: Jan 20, 2015
Copyright: © 2015 Nicolini C, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

We present a useful approach towards for biomarkers identification in an innovative self-assembling protein microarray based on “Nucleic Acid Programmable Protein Array” (NAPPA) and SNAP tag coupled to E.coli cell free expression system. This approach prove capable to resolve the “background” problem associated to the above label free detection system for the identification of proteins and of protein-protein interaction in humans that could become used in clinical practice.

Keywords: Mass Spectrometry; Electrophoresis; Protein

Introduction

In the last decade, Mass Spectrometry has played a key role in the advance of proteomics [1,2]. As research moves toward more sophisticated systems, it is urgent to develop protein analysis and identification techniques to meet the high-throughput demand [3-8]. The integration of microarrays with MS has generated a powerful new tool to deal with the problems in this area [9]. The flight time between the laser striking the array surface and the molecules reaching the detector at the end of the tube depend on the m/z of the proteins, thus enabling the system to accurately determine the mass of the protein species present in the sample [1,10]. One reported successful example is the ProteinChip® System of Ciphergen Biosystems Inc consisting in a SELDI-TOF-MS instrument equipped with a pulsed UV nitrogen laser source. Upon laser activation, the proteins at the array surface are desorbed and ionized, and subsequently accelerated by an electric field down the flight-tube, before reaching the detector. The patent (i.e. EU Patent No. 1 354 203) describes using mass spectrometry to detect certain protein biomarkers that are present in patients with bladder cancer versus patients who do not have bladder cancer. The high specificity of MS means that the signals of minute proteins or peptides that are undetectable using traditional techniques can be measured. As a result, SELDI-TOF MS has been applied to the screening of tumor biomarkers such as ovarian cancer, urinary bladder cancer, lung cancer, prostate cancer, colon cancer, breast cancer and liver cancer. Another example of the detection of a biomarker was the identification of CD8 cell anti-HIV factor (CAF). It has been known for more than a decade that certain HIV-1-infected individuals who are immunologically stable secrete a soluble factor, CAF, which suppresses HIV-1 replication. Although considerable work has been done, their identity was still obscure. Zhang et al. used the SELDI technique to discover a cluster of small proteins that were secreted when CD8 T cells from longterm non-progressors with HIV-1 infection were stimulated [11]. Although the SELDI protein chip has many advantages, including simple operational procedures, speed, high sensitivity and abundant information, it faces several challenges, including the normalization of sample collection and experimental procedures, identification and verification of biomarkers efficiently, and proper interpretation of sophisticated SELDI-MS data. In addition, most proteins in serum have a very low concentration and are difficult to detected by the SELDI technique. This may require pre-enrichment or separation using beads, LC and electrophoresis [12].

In previous researches [1,13,14] we carried out feasibility studies of MALDI-TOF MS analysis of different kind of Nucleic Acid Programmable Protein Array (NAPPA). The NAPPA method allows for functional proteins to be synthesized in situ directly from printed cDNAs just in time for assay. The use of purified proteins was substituted with the use of cDNAs encoding the target proteins for the microarray. In our research we employed two different mass spectrometry (MS) techniques, the Matrix Assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) MS and Liquid Chromatography-Electrospray Ionization MS (LC-ESI-MS). The last goal of our research is to develop a standardized analysis procedure, able to analyze the protein-protein interactions occurred on NAPPA array in a label free manner.

In the present manuscript we present the data obtained by a bioinformatics analysis of MALDI-TOF MS data carried out utilizing a specific “PURE system” database. Taking advantage of the full characterization of PURE Express system, starting from the list of its components we constructed a database of the entire triptych fragment belonging to PURE system molecules. Using SpADS algorithms we subtracted from our experimental mass lists the background peaks belonging to PURE system molecules. Once this procedure will prove successful serum proteomic profiles and emerging protein-protein interactions computed from MALDI-TOF of NAPPA SNAP arrays in association with QCM_D nanoconductimetry [7,8] could be measured by MALDI TOF MS along with classification tree established via our software in order to help us to provide a more accurate approach for diagnosis and clinical staging of cancers.
Materials and Methods

For all concerns production and expression of NAPPA and MS analysis refer to [6].

NAPPA SNAP

We analyzed different kind of NAPPA, in the last improved version the proteins were synthesized with the addition of a SNAP tag – therefore we named SNAP_NAPPA this kind of array - and translated using a reconstituted Escherichia coli coupled cell-free expression system. The addition of a SNAP tag to each protein enabled its capture to the array through an anti-SNAP antibody printed simultaneously with the expression plasmid. SNAP tag is a 20 kDa mutant of the DNA repair protein O6-alkylguanine-DNA alkyltransferase that reacts specifically and rapidly with benzyguanine (BG) derivatives, leading to irreversible covalent labeling of the SNAP tag. SNAP tag has a number of features that make it ideal for a variety of applications in protein labeling, in particular its substrates are chemically inert towards other proteins, avoiding non-specific labeling in cellular applications. Moreover also the chemistry and the printing of the NAPPA have been improved [15]. The MS samples are obtained from SNAP-NAPPA spots printed on gold coated glass slides in higher density, in order to obtain an amount of protein appropriate for MS analysis. The spots of 300 microns were printed in 12 boxes, each box with 100 identical spots. The sample genes immobilized used as test cases were p53_Human (Cellular tumor antigen p53); CDK2_Human (Cyclin-dependent kinase), 2;Src_Human-SH2 (the SH2 domain of Proto-oncogene tyrosine-protein kinase), PTP11 (Human-SH2, the SH2 domain of Tyrosine-protein phosphatase non-receptor type 11).

"PURE system" database construction

To reduce the sample complexity (i.e. the amount of biological data due to NAPPA chemistry and to the expression system) the in vitro translation-transcription (IVTT) system we used was from E.coli. The PURE system represents an important step towards a totally defined in vitro transcription/translation system, thus avoiding the "black box" nature of the cell extract. The immediate advantage is the significantly reduced level of all contaminating activities. The PURE system has the capacity for a yield of more than 100 µg/ml is today exclusively licensed to New England Biolabs (Ipswich, MA, USA) under the trade-name "PURExpress" [16]. Moreover the E.coli IVTT lysate is totally characterized, which could be a fundamental advantage for the subsequent analysis of the results.

The base to realize the "PURE system" database was the full knowledge of PURE EXPRESS composition (reported in Table 1). Through Expasy databank (www.expasy.org) search we identified the peptide sequences for each component. These sequences were in silico trypsin digested by means of the software Sequence Editor included into the Biotools package. Hereafter the concentrations of the components used in the PURE system [17].

<table>
<thead>
<tr>
<th>EF-Tu</th>
<th>IleRS</th>
<th>ProRS</th>
<th>Creatine kinase</th>
</tr>
</thead>
<tbody>
<tr>
<td>EF-Ts</td>
<td>LeuRS</td>
<td>ThrRS</td>
<td>Nucleotide diphosphate. Kinase</td>
</tr>
<tr>
<td>EF-G</td>
<td>MetRS</td>
<td>SerRS</td>
<td>Myokinase</td>
</tr>
<tr>
<td>RF1</td>
<td>ValRS</td>
<td>LysRS</td>
<td>Inorganic pyrophosphatase</td>
</tr>
<tr>
<td>RF2</td>
<td>GluRS</td>
<td>AsnRS</td>
<td>T7 RNA polymerase</td>
</tr>
</tbody>
</table>

II_ribosomal proteins

| 30S ribosomal subunit protein S1 | 50S ribosomal protein L17 | 50S ribosomal protein L5 | 30S ribosomal protein S17 |
| 30S ribosomal subunit protein S2 | 50S ribosomal protein L18 | 50S ribosomal protein L6 | 30S ribosomal subunit S18 |
| 30S ribosomal subunit protein S3 | 50S ribosomal protein L19 | 50S ribosomal protein L7 | 30S ribosomal subunit S19 |
| 30S ribosomal subunit protein S4 | 50S ribosomal protein L20 | 50S ribosomal protein L7/12 | 30S ribosomal subunit S20 |
| 30S ribosomal subunit protein S5 | 50S ribosomal protein L21 | 50S ribosomal protein L9 | 30S ribosomal subunit S21 |
| 30S ribosomal subunit protein S6 | 50S ribosomal protein L22 | 50S ribosomal protein L10 | 30S ribosomal subunit S22 |
| 30S ribosomal subunit protein S7 | 50S ribosomal protein L23 | 50S ribosomal protein L11 | 30S ribosomal subunit S23 |
| 30S ribosomal subunit protein S8 | 50S ribosomal protein L24 | 50S ribosomal protein L12 | 30S ribosomal subunit S24 |
| 30S ribosomal subunit protein S9 | 50S ribosomal protein L25 | 50S ribosomal protein L13 | 30S ribosomal subunit S25 |
| 30S ribosomal subunit protein S10 | 50S ribosomal protein L26 | 50S ribosomal protein L14 | 30S ribosomal subunit S26 |
| 30S ribosomal subunit protein S11 | 50S ribosomal protein L27 | 50S ribosomal protein L15 | 30S ribosomal subunit S27 |
| 30S ribosomal subunit protein S12 | 50S ribosomal protein L28 | 50S ribosomal protein L16 | 30S ribosomal subunit S28 |

III_ribosomal RNAs

| 23S Rrna | 5S rRNA | 16S rRNA |

I_recombinant proteins

| IF1 | RF3 | GinRS | AspS |
| IF2 | RRF | TrpRS | AlaRS |
| IF3 | ArgRS | TyrRS | GlyRS |
| Methionyl-tRNA formyltransferase | CysRS | HisRS | PheRS α2β2 |

J Mol Biomark Diagn
ISSN:2155-9929 JMBD, an open access Journal
Volume 6 • Issue 1 • 1000214
The PURE system includes 32 components that we purified individually: IF1, IF2, IF3, EF-G, EF-Tu, EF-Ts, RF1, RF2, and RF3, as well as RRF for termination. However, RF2 is not required for the translation of genes terminating with the codons UAG or UAA. The PURE system represents a step towards a totally defined in vitro transcription/translation system, thus avoiding the “black box” nature of the cell extract. The immediate advantage is the significantly reduced level of all contaminating activities and The E.coli IVTT with spades was used for the subtraction of the Master Mix spectrum from p53 and ptp spectra respectively [13]. The options used for the preprocessing of these latter two spectra were a binning window of 100 and peak extraction. No Region of Interest were selected, i.e. the whole range of the spectra were used. Finally, before the background subtraction, a peak alignment was performed.

SpADS an R implementation of preprocessing algorithms for data reduction and noise suppression was used in order to filter results from background noise i.e. master mix MS spectrum. Moreover, this latter was used coupled to and R implementation of the K Means clustering.

**Results**

The goal is to develop a standardize procedure to identify biomarkers in clinical setting and to analyze the protein-protein interaction occurred on NAPPA array using Matrix Assisted Laser Desorption Ionization Time-of-Flight (MALDI-TOF) Bruker Autoflex. We employ in the process "Protein synthesis Using Recombinant Elements" (PURE) system which due to its high complexity needs ad hoc bioinformatic tools to be analysed. The PURE system represents a step towards a totally defined in vitro transcription/translation system, thus avoiding the “black box” nature of the cell extract. The immediate advantage is the significantly reduced level of all contaminating activities and The E.coli IVTT with respect to the RRL or human lysate, which is totally characterized and thereby represents an advantage for the subsequent MS analysis of the results. The presence of “background” molecules, in fact, represents the main obstacle to these MS data interpretation. For this latter reason SpADS: An R Script for Mass Spectrometry Data Preprocessing before Data Mining an ad hoc script was implemented. SpADS provides useful preprocessing functions such binning, peak extractions, spectra background subtraction and dataset managing. Moreover, in its final version, it is able to perform peak recognition.
and amplitude independent subtraction functions were implemented [13].

Figure 2: SpADS and Clustering solution for a specimen of 56 protein samples of raw data. Only binning preprocessing function was performed before cluster analysis run on the ROI 1000/1200.

Results are showed in Figures 3-6. To reduce the sample complexity (i.e. the amount of biological material due to NAPPA chemistry and to the expression system) the in vitro translation-transcription (IVTT) system we used was from E.coli. The PURE system represents an important step towards a totally defined in vitro transcription/translation system, thus avoiding the “black box” nature of the cell extract. The immediate advantage is the significantly reduced level of all contaminating activities. The PURE system has the capacity for a yield of more than 100 µg/ml is today exclusively licensed to New England Biolabs (Ipswich, MA, USA) under the trade-name “PURExpress” [17] Moreover the E.coli IVTT lysate is totally characterized, which could be a fundamental advantage for the subsequent analysis of the results.

The base to realize the “PURE system” database was the full knowledge of PURE EXPRESS composition. Through Expsy databank (www.expasy.org) search we identified the peptide sequences for each component. These sequences were in silico trypsin digested by means of the software Sequence Editor included into the Biotools package. Hereafter the concentrations of the components used in the PURE system [17].

The proteins immobilized on the SNAP are synthesized with a SNAP tag and a FLAG tag that could also contribute to the difficulty in matching spectra with databases that are based on tryptic digests of natural proteins. It was then useful to consider strategies that compensate for this.

We have, then, modified the sequence of our proteins, adding the tag sequences (the full protein sequences were obtained from NEB). We used this modified sequence to perform a new fingerprint: the theoretical mass lists of the chimeras after trypsin digestion by means of the software SequenceEditor included into the Biotools package. We matched the experimental mass lists with these theoretical mass lists.

Figure 3: Reconstructed MS spectrum obtained adding five different theoretical mass lists of PURE express components (reported in the legend)

Figure 4: Reconstructed MS spectrum obtained subtracting from CDK2 experimental mas list the PURE express compnents theoretical mass lists

In Figure 1 it is reported a theoretical mass spectrum, reconstructed starting from the theoretical mass list of different PURE system components (as reported in the figure legend), after trypsin digestion, by means of Microsoft Excel software. It is evident the high complexity of such kind of analysis without the aid of a specific software.

In Figures 2 and 3 are reported the experimental mass spectra of Cdk2 and p53 trypsic digested samples obtained by Microsoft Excel software after the subtraction of the theoretical mass lists of tryptic fragments of all the PURE systems components. For PTPN11 SH2 and SRC SH2 no peak remained after the background subtraction that is probably due to a lower level of expression of these proteins. In summary out of the total 140 lists summarized in Tables 1 and 2 only 5 different theoretical mass lists (reported in the side legend) of the PURE express components bacterial lysate are shown in Figure 3.
Figure 5: Reconstructed MS spectrum obtained subtracting from p53 human experimental mas list the PURE express components theoretical mass lists.

Figure 6: The bottom image is produced subtracting from p53 spectra the Master mix spectra properly aligned. The software cannot instead produce significant results automatically subtracting the bacterial lysate from the NAPPA spectra.

Table 2: Peptide Sequence of the component I of PURE Express

And even these few are difficult to distinguish. Figures 4 and 5 represent the reconstructed MS spectra obtained subtracting respectively from CDK2 and p53 experimental mas lists all the peaks of the PURE express components theoretical mass lists, after a very long work utilizing excel. For the experimental mass lists of SRC e PTP genes samples nothing remains visible (not shown). A satisfactory result is that some peaks are still present in half of our sample genes, suggesting that with the aid of ad hoc software this kind of analysis will improve significantly the end results. Encouragingly Figure 6 show in the bottom image a similarly good result is obtained by subtracting from p53 spectra the experimental Master mix spectra when properly aligned. The software cannot instead produce significant results automatically subtracting the bacterial lysate Master Mix from the NAPPA spectra.

Conclusions

In the present manuscript we have successfully carried out a proof of principles which however need further optimization of the experimental layout in progress. Recent development the monitoring of gene-gene [6,7,14,18,19] and protein-protein [20] interactions in SNAP NAPPA microarray by QMC_D nanoconductimetry [8], Mass Spectrometry [10], Anodic Porous Allumina, [21] and Bioinformatics [14] open new avenues in functional proteomics overcoming the critical limits of fluorescence clinical studies using Nucleic Acid Programmable Protein Arrays or similar [22]. It appears thereby of fundamental importance to combined Nanogenomics and Nanoproteomics to warrant significant advancements in clinical
research in general and in cancer treatment in particular. Our main pertinent findings characterizing several model system and several nanotechnologies support these conclusions and progress achieved in the improvement of automated label free biomarkers detection in NAPPA SNAP microarrays by Mass Spectrometry and subsequent sophisticated data acquisition and processing.

Acknowledgments

This research was supported by a MIUR (Ministry of University and Research of Italy) grant for Funzionamento to Fondazione ELBA Nicolini.

References