

1 **SalivaPRINT Toolkit – Protein profile evaluation and phenotype**
2 **stratification**

3 **Igor Cruz¹, Eduardo Esteves¹, Mónica Fernandes¹, Nuno Rosa¹, Maria**
4 **José Correia¹, Joel P. Arrais², Marlene Barros^{1*}**

5

6 ¹ Universidade Católica Portuguesa, Center for Interdisciplinary Research in
7 Health (CIIS), Institute of Health Sciences (ICS), Viseu, Portugal

8 ² Department of Informatics Engineering (DEI), Centre for Informatics and
9 Systems of the University of Coimbra (CISUC), University of Coimbra, Coimbra,
10 Portugal

11 * Corresponding author

12 Address correspondence to: Marlene Barros, PhD, Director of Center for
13 Interdisciplinary Research in Health (CIIS), Senior Scientist at SalivaTec
14 Universidade Católica Portuguesa, Estrada da Circunvalação 3504-505 Viseu –
15 Portugal; Tel. +351232430200 - Fax +351232428344
16 email: mbarros@viseu.ucp.pt

17

18 **Abstract**

19 The value of the molecular information obtained from saliva is dependent on the
20 use of *in vitro* and *in silico* techniques. The main proteins of saliva when
21 separated by capillary electrophoresis enable the establishment of individual
22 profiles with characteristic patterns reflecting each individual phenotype. Different
23 physiological or pathological conditions may be identified by specific protein

24 profiles. The association of each profile to the particular protein composition
25 provides clues as to which biological processes are compromised in each
26 situation. Patient stratification according to different phenotypes often within a
27 particular disease spectrum is especially important for the management of
28 individuals carrying multiple diseases and requiring personalized interventions.
29 In this work we present the SalivaPRINT Toolkit, which enables the analysis of
30 protein profile patterns and patient phenotyping. Additionally, the SalivaPRINT
31 Toolkit allows the identification of molecular weight ranges altered in a particular
32 condition and therefore potentially involved in the underlying dysregulated
33 mechanisms. This tutorial introduces the use of the SalivaPRINT Toolkit
34 command line interface (<https://github.com/salivatec/SalivaPRINT>) as an
35 independent tool for electrophoretic protein profile evaluation. It provides a
36 detailed overview of its functionalities, illustrated by the application to the analysis
37 of profiles obtained from a healthy population *versus* a population affected with
38 inflammatory conditions.

39 **Keywords**

40 Protein profiling; Protein pattern recognition; Protein phenotypes

41

42 **1. Introduction**

43 In the age of precision medicine, diagnostics are based on the use of big data
44 from genomic, proteomic and transcriptomic research. These techniques enable
45 the establishment of molecular phenomes associated with different functional
46 profiles which characterize the phenotypes of individuals sharing similar
47 conditions and may direct a personalized intervention.

48 Omics results have revealed information on molecules which are dysregulated

49 in certain health and disease situations. This information is stored in several
50 public databases [1–4].

51 Saliva is a fluid increasingly used in diagnostics [5]. Several techniques have
52 been used to acquire molecular information from this fluid. Such information is
53 available in several public databases such as OralOme [3,6] or SalivaOmics [7].
54 One of the techniques used to characterize the main protein content of saliva is
55 electrophoresis, and capillary electrophoresis is one of the most sensitive
56 variants. Despite the wide availability of capillary electrophoresis-based
57 techniques, the challenge remains in the exploration of the technique's full power.
58 In particular, the fact that the currently available tools for result analysis require
59 manual and visual inspection of the profiles and are not amenable to high
60 throughput result analysis, has created a bottleneck in the generation of powerful
61 analysis of the results from large number of profiles such as those generated in
62 large population studies.

63 Few studies have been developed to surpass these problems mainly in the
64 analysis of nucleic acid results [8,9] but also for total protein profiles [10,11].

65 In spite of the existence of studies to recognize patterns of capillary
66 electrophoresis profiles [12] there is, to our knowledge, no approach developed
67 and applied to the use of total protein profiles of complex samples for patient
68 stratification or sample quality control.

69 The possibility of establishing protein profile patterns corresponding to specific
70 clinical situations is an opportunity for the development of new diagnostics
71 strategies essential for the analysis of large samples characteristic of population
72 wide and large epidemiologic studies.

73 The Experion™ automated electrophoresis system [12] (from Bio-Rad

74 Laboratories, USA) was used to provide the data in the example presented in this
75 tutorial. This system integrates protein analysis into a single process in which
76 protein separation, staining, band detection and quantitation are automatically
77 executed and produces protein profiles in about 30 minutes (10 samples) through
78 an automated process.

79 By performing capillary electrophoresis it is possible to obtain a protein profile
80 of the sample within the molecular weights (MW) in the range of 10–260
81 kiloDaltons (kDa) while separating and detecting protein concentrations in the
82 2.5–2000 ng/mL range [12].

83 The system software is responsible for plotting the fluorescence index as a
84 function of migration time to produce an electropherogram. A virtual gel image is
85 generated from the electropherogram data. Proteins bands or peaks are
86 identified by migration time relative to the known MW markers.

87 After running the samples, relevant peak heights and density of protein bands
88 are calculated by the software and the output is exported in a file containing
89 multiple information such as MW, peak height, protein concentration, and total
90 sample concentration among others. This information can be used with data
91 analysis techniques in order to characterize each individual and/or the population
92 to which it belongs.

93 Capillary electrophoresis technology has been used efficiently to detect
94 *Listeria monocytogenes* in foods [13] and to measure ovarian cancer or cancer-
95 related proteins biomarkers in serum [14], however the methodology followed for
96 result processing was to manually select individuals and check which molecular
97 weights were different according to the individual's conditions.

98 The development of solutions for automatic analysis of the results produced

99 by capillary electrophoresis technology, to obtain typical profiles or molecular
100 weight ranges, revealing altered protein quantities, are a first approach to
101 evaluate the functional status of each individual. These solutions are also useful
102 for the identification of the molecular weight ranges in which there are
103 dysregulated proteins associated to specific pathologies or phenotypes and
104 therefore may be used for diagnosis or stratification.

105 SalivaPRINT Toolkit provides a set of functionalities to analyze the output data
106 provided by capillary electrophoresis techniques. This tool can be widely applied
107 for the analysis of data from protein separation techniques resulting in an output
108 of migration/molecular weight data and respective protein quantification in each
109 sample.

110

111 **2. The SalivaPRINT Toolkit command line tools**

112

113 **a. Installation**

114

115 SalivaPRINT Toolkit command line tools are written in Python and work on
116 Windows, macOS and Unix. Python 3.0 (<https://www.python.org/downloads/>) is
117 required along with the modules numpy (<http://www.numpy.org/>), scipy
118 (<https://www.scipy.org/>), configparser
119 (<https://docs.python.org/2/library/configparser.html>) and matplotlib
120 (<https://matplotlib.org/>).

121 After successful installation of Python and the required libraries for running the
122 program, the user should decompress the file *salivaprint.zip* to a new directory
123 and use the *salivaprint.py* as a normal program passing commands as

124 arguments. In order to check if everything is working properly, the command
125 *salivaprint.py -v* should print the version number as follows.

```
$ python salivaprint.py -v  
numpy version 1.12.0  
scipy version 0.19.0rc2  
matplotlib version 2.0.0  
SalivaPRINT version 0.1
```

126

127

b. Available commands

128

129 SalivaPRINT Toolkit is a command line tool, which allows data extraction and
130 analysis from capillary electrophoresis systems output files.

131 The functionalities available allow the construction of a matrix of molecular
132 weights from an output file provided by Experion™ systems, which can then be
133 used with data analysis and machine learning tools in order to find similarities
134 between individuals and/or populations. By implementing a naïve Bayes
135 classification algorithm, a probabilistic classifier based on the application of the
136 Bayes' theorem with strong independence assumptions between features, it
137 becomes possible to achieve an overview of important features for the
138 stratification of the individuals in study.

139 SalivaPRINT Toolkit available commands can be checked anytime by using -
140 **h** as argument. The following commands are currently implemented (version 0.1):

141

142 **-v**: Displays the program and required libraries version;

143 **-h**: Displays the help menu. Lists the available commands;

144 **-build** output_file: Builds a new molecular feature matrix from capillary
145 electrophoresis output files using *config.cfg* as the configurations file;

146 **-view** input_file: Shows a visual representation of the dataset previously
147 built using the **-build** flag;
148 **-learn** input_file output_file: Builds a classifier from input_file dataset.
149 Uses the name given as output_file for saving the created classifier;
150 **-classify** classifier_file dataset: Classifies the dataset using the previously
151 trained classifier.

152

153 **c. Dataset preparation**

154

155 The main data file accepted by SalivaPRINT Toolkit is composed by a Comma
156 Separated File (**CSV**) file with peak information collected with Experion™ (or
157 other equivalent system) in the format: Sample, Molecular Weight, Protein
158 Concentration, Sample Concentration without header information.

159 An example is shown below.

160

```
Sample1,9.57,43.86,786.87  
Sample2,12.89,12.71,786.87  
Sample3,16.11,124.98,786.87  
Sample4,27.70,43.29,786.87  
Sample5,9.64,42.46,721.85  
...
```

161

162

163 Linux command line tools provides an easy way to prepare the Experion™
164 output files as datasets which can be used with SalivaPRINT Toolkit. Assure the
165 use of a **CSV** format files containing the data encoded to **UTF-8** with Unix Line
166 Feed (**LF**) as line break special characters. Note that it is important to use this file
167 encoding since the **awk** language for processing text, available on the standard

168 Linux bash, may fail to correctly recognize columns if the file encoding is not
169 correctly set.

170 Using **awk** is a fast option to select the correct columns for creating the dataset
171 file. The following command selects rows 7,10,13 and 17 from all the data
172 available. Note that these row positions (7,10,13,17) correspond to the columns
173 which provide information as sample name, MW, concentration and sample
174 concentration in the standard output file, and are the ones we need in order to
175 use SalivaPRINT Toolkit .

```
176          awk -F',' '{print $7,$10,$13,$17}' output_experion.csv >  
177 dataset.csv
```

178

179 **d. Configurations file**

180

181 Config.cfg is the file that contains all the configurations necessary for the
182 program to run. In order to extract data from the original MW from the capillary
183 electrophoresis output file the following configurations are necessary.

184 **MIN_MOL_WEIGHT** – (Default 9) Minimum molecular weight, defined in kDa
185 to consider while extracting data from the input dataset file.

186 **MAX_MOL_WEIGHT** – (Default 120) Maximum molecular weight, defined in
187 kDa to consider while extracting data from the input dataset file.

188 **N_SLICES** – (Default 120) Number of slices to consider from the
189 MIN_MOL_WEIGHT to MAX_MOL_WEIGHT.

190 **DATASET** - Input file containing all the capillary electrophoresis molecular
191 weights at which protein concentration peaks occur.

192 **CONTROL** – A list of healthy individuals, or control individuals, present in the
193 DATASET file. It should contain the sample IDs as found in the DATASET one
194 by each line. Ideally, it should have the same length of STUDY list for generating
195 a balanced classifier.

196 **STUDY** - A list of unhealthy, or disease carrier individuals, present in the
197 DATASET file. It should contain the sample IDs as found in the DATASET one
198 by each line. Ideally, it should have the same length of CONTROL list for
199 generating a balanced classifier.

200

201 **3. Case study: What can we learn from patients with inflammatory** 202 **conditions?**

203

204 In order to build this tutorial, 184 salivary electrophoretic profiles from
205 Experion™ automated electrophoresis system were used. The data was split into
206 two classes regarding the health status of individuals. The healthy population was
207 composed of 92 individuals without acute or chronic inflammation, as far as could
208 be discerned from the clinical history, ranging from 18 to 89 years of age
209 (average: 23.7, standard deviation: 9.4). The unhealthy population was
210 represented by 92 individuals, ranging from 7 to 84 years of age (average: 39.4,
211 standard deviation: 25.3). These individuals presented a broad spectrum of
212 diseases, from oral problems such as gingivitis, to whole systemic and chronic
213 diseases as diabetes or celiac disease, all related to an underlying inflammatory

214 condition.

215

216 **a) Preparing the dataset**

217

218 For this part of the tutorial, the saliva protein profiles from 164 individuals (82
219 healthy and 82 inflammatory), were used. Considering that we have two output
220 files from SalivaPRINT Toolkit, one for patients with inflammation and one
221 without, we can process them using the following commands:

```
# Copies the files content, excluding headers, to the new file.  
tail -n +2 -q inflammatory_samples.csv >> experion_output.csv  
tail -n +2 -q healthy_samples.csv >> experion_output.csv  
# Removes unnecessary data columns.  
awk -F',' '{print $7,$10,$13,$17}' experion_output.csv > dataset.csv
```

222

223 After this procedure, dataset.csv should have the format shown in 2.c) and
224 should be is ready to be used with SalivaPRINT Toolkit.

225

226 **b) Building the Molecular Weight Matrix**

227

228 First, it is necessary to properly set the configurations file. Using a minimum
229 MW of 9 kDa and a maximum MW of 120 kDa with 120 slices we will get a
230 description of each individual protein profile. Experion™ does not account for MW
231 below 10kDa (~9.5) and identifications with MW above 120 kDa since these
232 larger MW are often protein aggregates easily formed in saliva [16]. Note that if
233 using a different sample it may be useful to include MWs above 120 kDa.

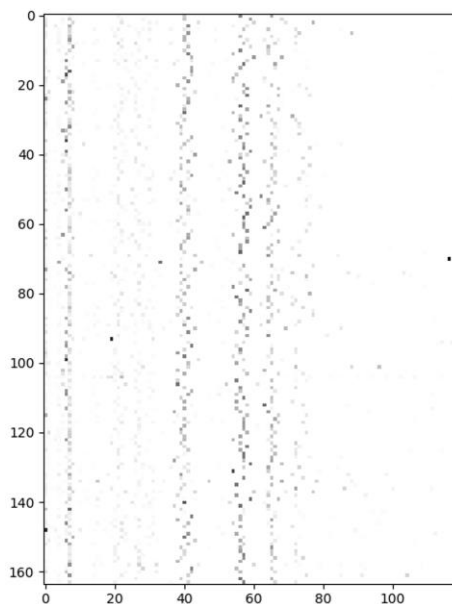
234 The configurations used are shown below.

```
[salivaprint_parameters]
MIN_MOL_WEIGHT: 9
MAX_MOL_WEIGHT: 120
NSLICES: 120
DATASET: dataset.csv
CONTROL: healthy_test.txt
STUDY: unhealthy_test.txt
```

235

236 The next step is to run **SalivaPRINT** Toolkit **-build matrix.csv** using the
237 standard configurations available in the configurations file. Make sure you build
238 two lists of individuals using the same IDs provided on the dataset file and edit
239 the config.cfg file to point to these files. One should list the healthy individuals
240 and the other the unhealthy. The program will then use the dataset in order to
241 build a matrix of relative concentration of protein per MW. This matrix represents
242 the presence of a ratio of protein.

243 By using the command **salivaprint.py -view matrix.csv** is possible to obtain
244 a visual representation of the matrix created.



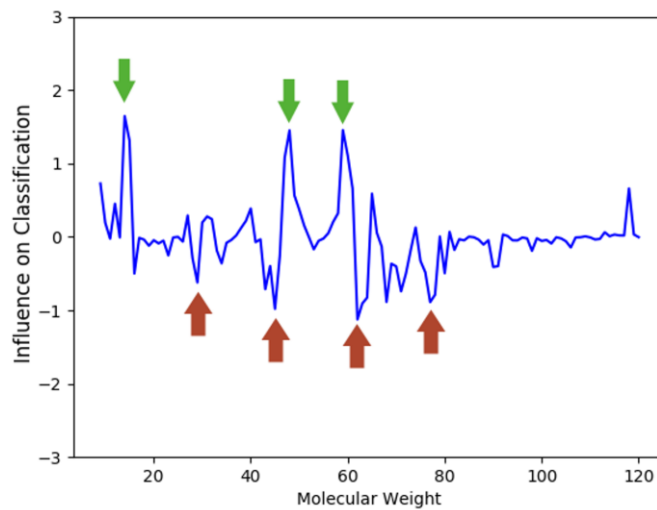
245

246 *Figure 1 - Graphical representation of the population. Each line represents one*
247 *individual and each column represents a small range of molecular weights (In*
248 *this case approximately 1kDa).*

249 c) Creating a Classifier

250

251 Using this matrix, which represents protein concentration per MW per
252 individual, it is possible to use SalivaPRINT Toolkit and create a classifier. The
253 command **salivaprint.py –learn matrix.csv classifier.pkl** will create a naïve
254 Bayes classifier with the examples provided in the matrix.csv file and save it with
255 the name classifier.pkl. When a Graphical User Interface (GUI) is available, it will
256 also show a graphical representation of the influence of each MW towards the
257 classification of samples according to condition state (healthy or inflammatory).
258 In Figure 2, we show the influence of molecular weights over the classification of
259 healthy individuals and individuals with inflammatory conditions obtained from the
260 dataset used on this tutorial. Y-axis values correspond to the influence of each
261 MW as learned from the naïve Bayes classifier. Negative values are associated
262 to the influence of a given MW over the population in study, in this case a
263 population with inflammatory conditions, while positive values are associated
264 towards the control population, in this case the healthy population.



265

266 *Figure 2 - Graphical representation of the MW influence towards classification*
 267 *of individuals. The green arrows point to MWs related to a tendency towards*
 268 *healthiness and the red arrows to MW related with inflammatory states.*

269 From this graphic representation it is possible to analyze the influence of
 270 different MW towards the classification of individuals given their protein profiles.
 271 Profiles containing some of the same MW as the positive values on Figure 2 are
 272 expected to be related to healthy individuals and profiles containing some of the
 273 same MWs as the negative values are expected to be related with individuals
 274 suffering from inflammatory conditions.

275

276 **d) Using a Classifier With a Different Dataset**

277

278 In table I a set of molecular weights and proteins within the ranges identified
 279 in the previous section is shown. In this table, the corresponding proteins are
 280 absent or present in different quantities. These MW ranges with the greatest
 281 variability in the proteins present enable through the identification of which

282 proteins are present (using Omics databases) and the potentially compromised
 283 molecular mechanisms. The potentially dysregulated proteins presented in each
 284 MW range were identified according to the data from Rosa *et. al*, 2016 [15].
 285 Proteins with molecular weights with a $\pm 8.56\%$ interval were considered since
 286 this is the largest variation in Experion™ efficiency as reported by the
 287 manufacturer [16].

288 *Table 1 – Proteins present in the MW ranges with greater influence in*
 289 *distinguishing protein profiles of healthy or inflammation challenged individuals.*

| <u>Molecular Weights</u> | <u>Proteins</u> |
|--------------------------|---|
| 14 – 15 kDa | P09228 Cystatin-SA (Cystatin-2) P01037 Cystatin-SN (Cystatin-SA-I) P01036 Cystatin-S (Cystatin-4) P01034 Cystatin-C (Cystatin-3) P07737 Profilin-1 Q01469 Fatty acid-binding protein P06702 Protein S100-A9 (Calgranulin-B) |
| 46 – 49 kDa | P52209 6-phosphogluconate dehydrogenase P80303 Nucleobindin-2 P01871 Ig mu chain C region Q8N4F0 BPI fold-containing family B member 2 P01009 Alpha-1-antitrypsin Q9UIV8 Serpin B13 P30740 Leukocyte elastase inhibitor (LEI) |
| 58 – 61 kDa | P14618 Pyruvate kinase PKM P04745 Alpha-amylase 1 P07237 Protein disulfide-isomerase (PDI) Q9UBG3 Cornulin P52209 6-phosphogluconate dehydrogenase |
| 28 – 29 kDa | P06870 Kallikrein-1 P31947 14-3-3 protein sigma Q96DR5 BPI fold-containing family A member 2 |
| 42 – 46 kDa | P80303 Nucleobindin-2 P01871 Ig mu chain C region Q8N4F0 BPI fold-containing family B member 2 |

| | |
|-------------|---|
| | P01009 Alpha-1-antitrypsin Q9UIV8 Serpin B13 P30740 Leukocyte elastase inhibitor (LEI) P04083 Annexin A1 P01876 Ig alpha-1 chain C region Q6P5S2 Protein LEG1 homolog |
| 62 – 64 kDa | P02768 Serum albumin P15311 Ezrin (Cytovillin) P14618 Pyruvate kinase PKM P04745 Alpha-amylase 1 P07237 Protein disulfide-isomerase (PDI) |
| 76 – 78 kDa | P01833 Polymeric immunoglobulin receptor (PIgR) P22079 Lactoperoxidase (LPO) P02788 Lactotransferrin (Lactoferrin) Q08188 Protein-glutamine gamma-glutamyltransferase E P02768 Serum albumin P15311 Ezrin (Cytovillin) |

290

291

e) Using SalivaPRINT Toolkit as a Classification Tool

292

293

Another functionality implemented in SalivaPRINT Toolkit is the possibility to

294

run the previously created classifier on an independent set of individuals. This

295

allows to verify if the classifier has correctly learned to differentiate the molecular

296

weights related with the two populations (when the expected output is known), as

297

well as providing a way to test if a particular individual is more similar to a

298

population or another.

299

In this step, a new set of individuals from the original dataset, not used in the

300

training of the algorithm, was used for testing the previously created classifier.

301

A list of 10 healthy individuals and 10 unhealthy individuals was created:

| healthy_test.txt (expected output: 0) | unhealthy_test.txt (expected output: 1) |
|---------------------------------------|---|
| Do1329 | Doo361 |
| Do1353 | Doo806 |
| Do1299 | Doo329 |
| Do1315 | Doo334 |
| Do1313 | Doo029 |
| Do1360 | Doo051 |
| Do1373 | Doo899 |
| Do1347 | Doo548 |
| Do1298 | Doo362 |
| Do1349 | Doo098 |

302

303 Next, the configurations file was adapted to create a testing dataset, note that

304 it must provide the same configurations as the ones used to extract the data,

305 which was used to create the classifier.

306

```
[salivaprint_parameters]
MIN_MOL_WEIGHT: 9
MAX_MOL_WEIGHT: 120
NSLICES: 120
DATASET: dataset.csv
CONTROL: healthy_test.txt
STUDY: unhealthy_test.txt
```

307

308 Then it is necessary to generate the test dataset, as follows:

309 **python salivaprint.py -build inflammation_test.csv**

310 And, finally, classify the test dataset:

```
python salivaprint.py -classify classifier.pkl
inflammation_test.csv
Do1329 0.379498254621 0.0
Do1353 0.326546267944 0.0
Do1299 0.435229022629 0.0
Do1315 0.190077337287 0.0
Do1313 0.378009272384 0.0
Do1360 0.571724373544 1.0
Do1373 0.166022320581 0.0
Do1347 0.374208781178 0.0
Do1298 0.235062846741 0.0
Do1349 0.517444020404 1.0
Doo361 0.507039962205 1.0
Doo806 0.662820518226 1.0
Doo329 0.642052560463 1.0
Doo334 0.646207130445 1.0
Doo029 0.531871564286 1.0
Doo051 0.507726735865 1.0
Doo899 0.708144099429 1.0
Doo548 0.58105270355 1.0
Doo362 0.624238965762 1.0
Doo098 0.513043859286 1.0
```

311

312 The values closer to zero are related with a tendency towards healthier states
313 and values closer to 1 are related with inflammatory states. As shown, using this
314 independent dataset, the classifier was able to correctly identify 18 out of 20
315 samples. Note that the misclassified examples occurred when the expected class
316 was 0 and present values closer to 0.5 than most of the samples where the
317 expected class was 1. This means that, despite misclassified, they are closer to
318 the threshold line, which splits the two classes.

319 It is also important to keep in mind that the inflammatory process is not a binary
320 classification problem in its origin; there are no absolute healthy or unhealthy
321 individuals from which the classifier can learn from. Thus, it is expected that small
322 changes in the classification threshold line (here considered to be 0.5) lead to
323 adaptations on the sensibility and specificity of the algorithm.

324

325 **4. Case study: Celiac patients a distinct phenotype within the** 326 **inflammatory process**

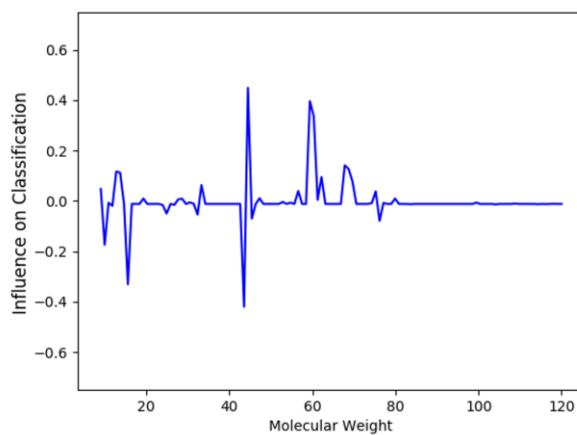
327

328 In this case study, a dataset of individuals diagnosed as celiac was used.
329 These individuals share a chronic inflammation status and therefore it is expected
330 that their salivary protein profile reflects the underlying functional dysregulation.

331 To test this hypothesis celiac patients were chosen according to time since
332 diagnostic and grouped in 1-5 years or more than 5 years since diagnostics. From
333 each of these groups the individuals presenting the most dysregulated protein
334 profiles were selected. This selection was based on complementary clinical data.

335 SalivaPRINT Toolkit commands were run following the steps above,
336 considering, the two groups of celiac patients. The goal was to find which MW is
337 important in the distinction of these groups.

338 The plot below represents the output of `salivaprint.py –learn` using this dataset.
339 The molecular differences found between the two groups were minimal occurring
340 on a small number of MW and with small values of influence (<0.5).



341

342 *Figure 3 - Graphical representation of the molecular weight related with*
343 *dysregulated proteins in the celiac groups.*

344 The small differences found between the two groups with different diagnostic
345 times are characterized by different profiles in the MW ranges presented in table
346 II. The potential dysregulated proteins are also presented in each MW range
347 according to the data from Rosa *et. al*, 2016 [15]. Proteins with molecular weights
348 with a $\pm 8.56\%$ interval were considered since this is the largest variation in
349 Experion™ efficiency as reported by the manufacturer [16].

350 *Table II – Proteins present in the MW ranges with greater influence in*
351 *distinguishing protein profiles of healthy or inflammation challenged individuals.*

| <u>Molecular Weights</u> | <u>Proteins</u> | |
|-------------------------------|--|--|
| 44 – 45 kDa (1-5 years) | P01871 Q8N4F0 P01009 Q9UIV8 P30740 P04083 | Ig mu chain C region BPI fold-containing family B member 2 Alpha-1-antitrypsin Serpin B13 Leukocyte elastase inhibitor (LEI) Annexin A1 |
| 59 – 60 kDa (1-5 years) | P14618 P04745 P07237 Q9UBG3 | Pyruvate kinase PKM Alpha-amylase 1 Protein disulfide-isomerase (PDI) Cornulin |
| 15 – 16 kDa (+5 years) | P27482 P12273 P02810 P09228 P01037 P01036 P01034 P07737 Q01469 | Calmodulin-like protein 3 Prolactin-inducible protein (PIP) Salivary acidic proline-rich phosphoprotein 1/2 Cystatin-SA (Cystatin-2) Cystatin-SN (Cystatin-SA-I) Cystatin-S (Cystatin-4) Cystatin-C (Cystatin-3) Profilin-1 Fatty acid-binding protein |
| 43 – 44 kDa (+5 years) | P01009 Q9UIV8 P30740 P04083 | Alpha-1-antitrypsin Serpin B13 Leukocyte elastase inhibitor (LEI) Annexin A1 |

352

353 **Conclusions**

354 SalivaPRINT Toolkit is a command line tool that uses machine learning to
355 analyze and learn from capillary electrophoresis data set experiments.

356 The analysis of individual protein profiles stratified by health condition has
357 enabled the proposal of which MW ranges and respective proteins are altered in
358 each group, leading to the inference of which molecular processes might be
359 compromised.

360 In this tutorial, two scenarios were selected to demonstrate the use of the
361 SalivaPRINT toolkit. First, a dataset composed of healthy individuals and
362 individuals suffering from inflammatory conditions. Second, a group of individuals
363 all with celiac disease, but stratified by date of diagnosis and treatment.

364 In both cases, the use of the proposed toolkit enabled the finding of protein
365 MWs ranges, which characterizes the protein phenotype of these individuals.

366 The true power of using SalivaPRINT Toolkit as protein profile analysis tool,
367 relies on the fact that the information of a large number of profiles is analyzed
368 simultaneously and large amounts of data are accounted for, enabling the
369 inference of which proteins may be involved with the underlying molecular
370 process compromised in a particular condition. In this way, the identification of
371 the protein profile patterns in saliva corresponding to different clinical situations,
372 or the existence of different patterns within the same pathology may constitute a
373 first approach to establish patient stratification according to the individual
374 molecular profile (phenotype).

375

376 **Acknowledgments**

377 This work was made possible by the support of Centre for Interdisciplinary
378 Research in Health through funds from the FCT (Portuguese national funding
379 agency for science, research and technology) granted to Unit 4279 -
380 UID/MULTI/4279/2016 and by the support of SalivaTec through + Centro -
381 Programa Operacional Regional do Centro under the Quadro de Referência
382 Estratégico Nacional (QREN) and through the Fundo Europeu de
383 Desenvolvimento Regional (FEDER) CENTRO-07-CT62-FEDER-005004.

384 Igor Cruz were supported by Fundação para a Ciência e Tecnologia (Grant No.
385 UID/MULTI/4279/2016.

386 **References**

387 [1] K.G. Becker, K.C. Barnes, T.J. Bright, S.A. Wang, The genetic association
388 database., *Nat. Genet.* 36 (2004) 431–2. doi:10.1038/ng0504-431.

389 [2] A. Kozomara, S. Griffiths-Jones, miRBase: annotating high confidence
390 microRNAs using deep sequencing data., *Nucleic Acids Res.* 42 (2014)
391 D68-73. doi:10.1093/nar/gkt1181.

392 [3] J.P. Arrais, N. Rosa, J. Melo, E.D. Coelho, D. Amaral, M.J. Correia, M.
393 Barros, J.L. Oliveira, OralCard: a bioinformatic tool for the study of oral
394 proteome., *Arch. Oral Biol.* 58 (2013) 762–72.
395 doi:10.1016/j.archoralbio.2012.12.012.

396 [4] M. Uhlén, L. Fagerberg, B.M. Hallström, C. Lindskog, P. Oksvold, A.
397 Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K.
398 Edlund, E. Lundberg, S. Navani, C.A.-K. Szigartyo, J. Odeberg, D.
399 Djureinovic, J.O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H.
400 Tegel, J. Mulder, J. Rockberg, P. Nilsson, J.M. Schwenk, M. Hamsten, K.
401 von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von
402 Heijne, J. Nielsen, F. Pontén, *Proteomics. Tissue-based map of the human*
403 *proteome.*, *Science.* 347 (2015) 1260419. doi:10.1126/science.1260419.

404 [5] F.M.L. Amado, R.P. Ferreira, R. Vitorino, One decade of salivary
405 proteomics: current approaches and outstanding challenges., *Clin.*
406 *Biochem.* 46 (2013) 506–17. doi:10.1016/j.clinbiochem.2012.10.024.

- 407 [6] N. Rosa, M.J. Correia, J.P. Arrais, P. Lopes, J. Melo, J.L. Oliveira, M.
408 Barros, From the salivary proteome to the OralOme: comprehensive
409 molecular oral biology., *Arch. Oral Biol.* 57 (2012) 853–64.
410 doi:10.1016/j.archoralbio.2011.12.010.
- 411 [7] D.T.W. Wong, Salivaomics., *J. Am. Dent. Assoc.* 143 (2012) 19S–24S.
412 <http://www.ncbi.nlm.nih.gov/pubmed/23034834>.
- 413 [8] S. Yoon, J. Kim, J. Hum, H. Kim, S. Park, W. Kladwang, R. Das, HiTRACE:
414 high-throughput robust analysis for capillary electrophoresis.,
415 *Bioinformatics.* 27 (2011) 1798–805. doi:10.1093/bioinformatics/btr277.
- 416 [9] S. Lee, H. Kim, S. Tian, T. Lee, S. Yoon, R. Das, Automated band
417 annotation for RNA structure probing experiments with numerous capillary
418 electrophoresis profiles., *Bioinformatics.* 31 (2015) 2808–15.
419 doi:10.1093/bioinformatics/btv282.
- 420 [10] M. Jonsson, J. Carlson, Computer-supported interpretation of protein
421 profiles after capillary electrophoresis., *Clin. Chem.* 48 (2002) 1084–93.
422 <http://www.ncbi.nlm.nih.gov/pubmed/12089178>.
- 423 [11] G.A. Ceballos, J.L. Paredes, L.F. Hernández, Pattern recognition in
424 capillary electrophoresis data using dynamic programming in the wavelet
425 domain, *Electrophoresis.* 29 (2008) 2828–2840.
426 doi:10.1002/elps.200700831.
- 427 [12] H. Laboratories, Electrophoresis System, (2000) 1–6. doi:10.1016/S0960-
428 9822(00)80094-9.
- 429 [13] E. Delibato, A. Gattuso, A. Minucci, B. Auricchio, D. De Medici, L. Toti, M.

- 430 Castagnola, E. Capoluongo, M.V. Gianfranceschi, PCR experion
431 automated electrophoresis system to detect *Listeria monocytogenes* in
432 foods., *J. Sep. Sci.* 32 (2009) 3817–21. doi:10.1002/jssc.200900166.
- 433 [14] J.H. Kim, Y.-W. Kim, I.-W. Kim, D.C. Park, Y.W. Kim, K.-H. Lee, C.K. Jang,
434 W.S. Ahn, Identification of candidate biomarkers using the Experion™
435 automated electrophoresis system in serum samples from ovarian cancer
436 patients., *Int. J. Oncol.* 42 (2013) 1257–62. doi:10.3892/ijo.2013.1803.
- 437 [15] N. Rosa, J. Marques, E. Esteves, M. Fernandes, V.M. Mendes, Â. Afonso,
438 S. Dias, J.P. Pereira, B. Manadas, M.J. Correia, M. Barros, Protein Quality
439 Assessment on Saliva Samples for Biobanking Purposes., *Biopreserv.*
440 *Biobank.* 14 (2016) 289–97. doi:10.1089/bio.2015.0054.
- 441 [16] K. Zhu, M. Nguyen, W. Strong, C. Whitman-guliaev, B. Laboratories, P.
442 Samples, Performance Comparison of the Experion™ Automated
443 Electrophoresis System and SDS-PAGE for Protein Analysis, *Methods.*
444 (n.d.) 0–5.