Anal Bioanal Chem (2012) 402:2521–2529
DOI 10.1007/s00216-011-5350-3

ORIGINAL PAPER

On the utility of predictive chromatography to complement mass spectrometry based intact protein identification

Marina L. Pridatchenko · Tatyana Yu. Perlova · Hisham Ben Hamidane · Anton A. Goloborodko · Irina A. Tarasova · Alexander V. Gorshkov · Victor V. Evreinov · Yury O. Tsybin · Mikhail V. Gorshkov

Received: 26 May 2011 / Revised: 22 July 2011 / Accepted: 19 August 2011 / Published online: 8 September 2011

Abstract The amino acid sequence determines the individual protein three-dimensional structure and its functioning in an organism. Therefore, "reading" a protein sequence and determining its changes due to mutations or post-translational modifications is one of the objectives of proteomic experiments. The commonly utilized approach is gradient high-performance liquid chromatography (HPLC) in combination with tandem mass spectrometry. While serving as a way to simplify the protein mixture, the liquid chromatography may be an additional analytical tool providing complementary information about the protein structure. Previous attempts to develop "predictive" HPLC for large biomacromolecules were limited by empirically derived equations based purely on the adsorption mechanisms of the retention and applicable to relatively small polypeptide molecules. A mechanism of the large biomacromolecule retention in reversed-phase gradient HPLC was described recently in thermodynamics terms by the analytical model of liquid chromatography at critical conditions (BioLCCC). In this work, we applied the BioLCCC model to predict retention of the intact proteins as well as their large proteolytic peptides separated under different HPLC conditions. The specific aim of these proof-of-principle studies was to demonstrate the feasibility of using "predictive" HPLC as a complementary tool to support the analysis of identified intact proteins in top-down, middle-down, and/or targeted selected reaction monitoring (SRM)-based proteomic experiments.

Keywords Proteins · Liquid chromatography · Retention time prediction · Mass spectrometry · Sequence variants

Introduction

Proteomics is a rapidly evolving technological platform that embraces multidisciplinary approaches to effectively characterize proteins on a large scale [1]. To expound the fundamental molecular mechanisms associated with cell and/or organelle function, global knowledge of the expressed proteomes under different environmental conditions is required, including protein sub-cellular localization, post-translational modifications (PTMs), protein sequence mutations, and interactions [2]. To function correctly, each cell depends on thousands of proteins to function in the right places at the right times. When a mutation alters a protein that plays a critical role in the body, pathological changes in an organism may result [3]. The challenges usually faced when studying the organism's proteome are the large variety of proteins, the high concentration diversity of different proteins in a sample, and the rapid changes in the cell's protein content during its lifetime. Development of analytical instrumentation and data analysis tools for species of biological importance is a rapidly growing and highly demanded area of method development efforts. The important building platform for this yet emerging field is "omics"

Published in the special issue *Analytical Sciences in Switzerland* with guest editors P. Dittrich, D. Günther, G. Hopfgartner, and R. Zenobi.

M. L. Pridatchenko · T. Yu. Perlova · A. A. Goloborodko · I. A. Tarasova · M. V. Gorshkov (✉)
Institute for Energy Problems of Chemical Physics,
Russian Academy of Sciences,
119991 Moscow, Russia
e-mail: mike.gorshkov@gmail.com

H. Ben Hamidane · Y. O. Tsybin (✉)
Ecole Polytechnique Federale de Lausanne,
1015 Lausanne, Switzerland
e-mail: yury.tsybin@epfl.ch

A. V. Gorshkov · V. V. Evreinov
Institute of Chemical Physics, Russian Academy of Sciences,
119991 Moscow, Russia

cascade which includes genomics, transcriptomics, proteomics, and metabolomics [4]. Despite a substantial progress in mass spectrometry-based protein characterization achieved recently in various application areas, its reliability and robustness remain a challenge [5]. A combination of liquid chromatography (LC) and tandem mass spectrometry (MS/MS) is a method of choice for the protein identification in a sample because of its high-throughput and sensitivity [6]. There are two main approaches that reduce sample complexity before a mass spectrometer. In the first one, the intact proteins or large protein fragments are separated externally to a mass spectrometer and then fragmented inside, followed by mass analysis of the protein fragments or the intact proteins (top-down proteomics [7]). On the contrary, in the MS-based "bottom-up" proteomics [8, 9], proteins are digested proteolytically and tandem mass analysis is performed on the proteolytic peptides separated by LC systems.

In either top-down or bottom-up approaches, LC plays a slave role as a separation tool that simplifies and purifies a mixture of either proteins, or the proteolytic peptides before the MS analysis. At the same time, protein sequence identification can be further improved by accurate prediction of the respective retention times (RT) during the chromatographic runs in both approaches, providing a second dimension in the protein search space. For example, matching between observed and predicted peptide retention times in the 2D LC-MS space can "filter out" a large number of misidentified peptides (false positives) found using MS/MS database searches [10–16]. To use LC data for the above purpose, a quantitative relation between retention times and the protein sequence has to be established. This relationship is achieved using an applicable sequence-dependent model for the RT prediction (the comprehensive reviews of peptide RT prediction models can be found in [17, 18]). The existing approaches to RT prediction are based on artificial neural networks (ANN) [19], various empirical models starting from the early Sanger's work [20], or other semi-empirical additive models [21, 22]. In another predictive model, the additive approach based on the amino acid composition was amended by the introduction of sequence-specific correction factors (SSRCalc model) [23]. Note that the most accurate RT prediction models, like the ANN or SSRCalc, require large datasets obtained from LC analysis of well-known peptides. Besides, the correct prediction of the elution order of peptides with re-arrangements of amino acid residues within known sequences has not been fully demonstrated yet. Most importantly, the above approaches have limited capabilities for prediction of retention times for large sequences such as intact proteins or the proteolytic peptides containing more than 40–50 amino acid residues. Due to the large training dataset requirement, the extension of "predictive" capabilities of empirically derived HPLC models for large macromolecules is a challenging and time-consuming task. In summary, there is no general analytical theory describing from the first principles the separation of large-size biomacromolecules such as proteins.

Recently, we have introduced phenomenological description of biomacromolecule separation called Liquid Chromatography of Biomacromolecules at Critical Conditions (BioLCCC) [24, 25]. The model relies on a number of phenomenological parameters, including the interaction energies of amino acid residues with the adsorbent surface, specific to particular separation conditions (ion-pairing reagents, type of the stationary phase and its surface chemistry [26], etc.). These parameters are directly measured using a set of well-defined sequences. After these parameters being determined, the model is able to predict the RT of a protein sequence, thus providing additional sequence identification tool, complementary to MS/MS. In spite of being fully analytical (contrary to empirical approaches) the model has demonstrated its ability to predict retention times for tryptic peptides with the determination coefficient of $R^2 \sim 0.87$ to 0.95. The main concept behind the model is based on the "random-walk" chain approximation introduced in chromatography by DiMarzio and Rubin [27] to describe the behavior of large chain-like macromolecular structures and later advanced into the comprehensive treatment of retention of synthetic polymers and oligomers as a concept of critical liquid chromatography (LCCC) [28]. Therefore, from the intrinsic nature of the BioLCCC model, it presents the opportunity to better, comparably with relatively short peptides, describe retention of large protein sequences, for which the "chain" approximation becomes rather bold and inaccurate.

In this work, we applied the BioLCCC model to predict retention times for intact proteins as well as large-size proteolytic peptides separated under various reversed-phase chromatographic conditions. To demonstrate its predictive capabilities we selected proteins with minor differences in amino acid sequences, termed here as "point mutations" or "sequence variants", as well as their digests obtained using pepsin-based proteolysis.

Experimental

LC-MS grade water and acetonitrile (ACN) were purchased from Sigma-Aldrich (Buchs, Switzerland). Trifluoroacetic acid (TFA) and formic acid (FA) were purchased from Fisher Scientific (Loughborough, UK) and Merck (Darmstadt, Germany), respectively. *Bovine* and *canine* cytochrome *c* (C3131 and C4013, respectively), and *porcine* pepsin (P700) were obtained from Sigma Chemical Company (Steiheim, Germany). *Equine* cytochrome *c* was purchased from Calbiochem (Darmstadt, Germany). The HPLC Janeiro

CNS quaternary pump (Thermo Scientific, Switzerland) used in this study was coupled to the electrospray ionization source of a linear ion trap mass spectrometer (LTQ XL from Thermo Scientific, Bremen, Germany). A Discovery BIO WidePore C18300 Å, 150×2.1 mm, 5 µm particle size column was purchased from Supelco (Bellefonte, USA). Proteins and their proteolytic peptides were separated using linear gradients (solvent A: water 95%, ACN 5%, 0.1% FA, solvent B: water 5%, ACN 95%, 0,1% FA) at room/ambient temperature for flow rates of 0.2 mL/min.

The in-house developed software package "Theoretical chromatograph-BioLCCC" was used to calculate retention times (RT) for proteins and peptides. The software is currently available for: (1) two types of ion-pairing reagents, TFA and FA; (2) different types of stationary phase chemistry including the polar group-embedded reversed-phase C18 AQ (the FA-based BioLCCC model was implemented for AQ phase only [28]); (3) several amino acid modifications: specifically, phosphorylation, methionine oxidation and cystein carboxyamidomethylated; and (4) different C- and N- terminal groups including C-terminal amidation and N-terminal acetylation. The open-source libraries for "Theoretical chromatograph-BioLCCC" are available on-line at http://theorchromo.ru for both FA- and TFA-based separations. In addition to the data generated in the present study, HPLC data for intact protein separations, available in the literature [29–31], were used to compare experimental and predicted retention times.

Cytochrome *c* proteins from different organisms were digested using pepsin according to standard procedures described elsewhere [32]. Digestion was performed with a substrate ratio (enzyme to protein) of 1:20. The resulting proteolytic peptides were identified from tandem mass spectra using Mascot search engine (v.2.1) through the SwissProt database (selected enzyme was "pepsin A"). Data-dependent MS/MS acquisitions (on the five most intense peaks observed in MS^1) were performed using preliminarily optimized tune file and a mass/charge range comprised between 300 and 2,000 *m*/*z*.

Result and discussion

Theoretical background of the BioLCCC model

The system of basic equations in BioLCCC model has been previously reported elsewhere [24, 25]. Briefly, it consists of the equation for the effective energy of interaction between the amino acid residue and the surface, $\varepsilon^{\mathrm{eff}}{}_{i}$ (phenomenological parameter of the model); the equation for the chromatographic distribution coefficient K_d that is resulted from DiMarzio–Rubin's "random-walk" model [27], the relationship between the solvent strength, ε_{AB},

Table 1 HPLC chromatographic systems and conditions from refs. 29–31 used in present study

№	column	flow rate, ml/min	"A" solvent	"B" solvent	gradient profile, (min;%B)	literature source
1	Aquapore RP 300 C8 220×4.6 mm	1.0	0.1%TFA in H_2O	0.1%TFA in ACN	(0;0) (100;100)	[30]
2	Astec RP C18 150×6.0	1.0	0.05%TFA in H_2O/CAN (90/10)	0.05%TFA in H_2O/ACN (40/60)	(0;0) (30;100)	[31]
3	Macrosphere 300 C4 100×4.6	1.0	0.15%TFA in H_2O	0.15%TFA in H_2O/ACN (5/95)	(0;25) (30;100)	[31]
4	Macrosphere 300, C8, 250×4.6	1.0	0.15%TFA in H_2O	0.15%TFA in H_2O/ACN (5/95)	(0;25) (30;100)	[31]
5	Vydac 228 TP (218) C18 250×4.6	1.5	0.25%TFA in H_2O	0.15%TFA in H_2O/ACN (30/70)	(0;5) (30;100)	[31]
6	Spherisorb 300A C18 250×4.6	1.5	0.1%TFA in H_2O	100%ACN	(0;20) (30;80)	[31]
7	Astec RP C18 150×5.0	1.0	0.05%TFA in H_2O/ACN (80/20)	0.05%TFA in H_2O/ACN (50/50)	(0;0) (20;100)	[31]
8	Kromasil C8 250×4.6	2.0	0.1%TFA in H_2O/ACN (90/10)	0.1%TFA in H_2O/ACN (10/90)	(0;0) (8;25) (20;70)	[31]
9	Vydac 219 TPCY C4 250×4.6	1.0	0.1%TFA in H_2O	0.1%TFA in H_2O/ACN (30/70)	(0;0) (50;50) (60;70) (67.5;100)	[31]
10	In-house C8 Lx4.6 mm	1.5	0.1%TFA in H_2O	0.1%TFA in ACN	(0;15) (30;60)	[29]

In all referenced works, the reversed phases with pore size of 300 Å were used. In the work by Koyama et al. [29] (10th line in the table) four columns with different lengths of 10, 35, 100, 250 mm were used

Table 2 Comparison of experimental and predicted retention times for intact proteins separated in [29–31]

Protein	N	MW, kDa	Average RT_{exp}, min	RT (BioLCCC), min	RT (SSRCalc [15]), min	RT (Additive models), min		
						[13]	[29]	[30]
Peptide S4	10	1	10.2	8.2	10.8	11.8	11.9	8.4
Bovine Insulin (B-chain)	36	3.7	15.4	16.9	14.8	12.5	12.4	12.9
Horse Cytochrome C	104	11.7	16.1	17.1	16.1	13.7	13.5	16.3
Bovine Ribonuclease A	124	13.7	14.0	15.1	13.2	13.9	13.8	16.8
Chicken lysozime	129	14.3	17.5	17.5	16.2	14.9	14.5	17.4
Whale myoglobin	152	17	20.1	20.5	19.5	16.2	15.7	18.3
Rabbit RsTnC	159	18	21.7	20.9	21.6	16.6	16.1	18.6
Bovine α-chymotrypsinogen	245	25.7	20.5	19.4	18.2	19.0	18.7	20.1
Rabbit RcTM	284	32.7	19.1	17.2	17.6	18.6	18.5	20.3
Ovalbumin (common turkey)	386	43	22.4	22.6	22.0	25.5	25.7	22.1
Yeast enolase	436	46.7	20.9	20.7	23.5	24.7	25.2	22.2
BSA	583	66.4	18.6	20.5	23.1	29.2	30.6	23.3

Experimental retention times were converted into the universal scale of absolute retention times corresponding to the HPLC protocols used for the system #5 (Table 1). Experimental RTs obtained for all HPLC systems and protocols were then averaged for each of the proteins. Predicted RTs were obtained for the system #5 HPLC conditions using five retention models: BioLCCC, SSRCalc, Retention Coefficient (RC) theory [21, 22], and RC theory by Sakamoto et al. with sequence length correction [38]

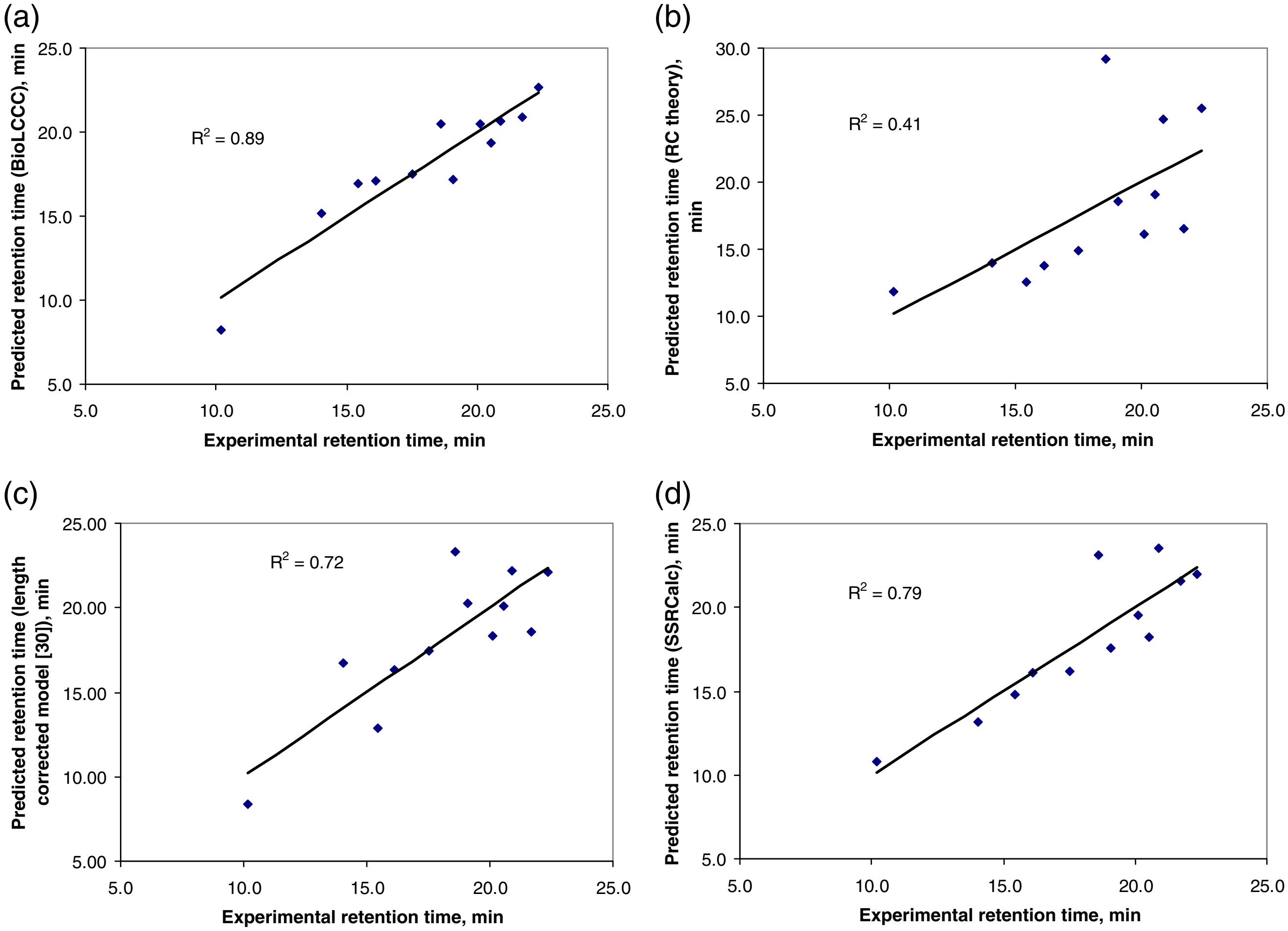

Fig. 1 Correlations between experimental and predicted retention times for 11 proteins under study for different retention models: **a** BioLCCC model; **b** additive model (RC theory [22]); **c** additive model with sequence length correction proposed in [38]; **d** SSRCalc model by Krokhin et al. [23]

obtained within Langmuir isotherm approximation and which is the free energy change for organic solvent molecules during its adsorption/desorption at the surface, and the equation for the retention volume V_R in the gradient LC:

$$\begin{aligned}
&\varepsilon_i^{eff}(N_B) = \alpha\left(X_i^0 - \varepsilon_{AB}(N_B)\right)\\
&K_d = \frac{1}{D} U^T \prod_{i=2}^{n} W\left(\varepsilon_i^{eff}\right) P_0\\
&\varepsilon_{AB}(N_B) = \varepsilon_A + \ln[1 - N_B(V) + N_B(V)\exp(\varepsilon_B - \varepsilon_A)]\\
&\int_0^{V_R - V_0} \frac{dV}{V_p \cdot K_d(V)} = 1
\end{aligned} \tag{1}$$

Here, the subscript i denotes an amino acid residue in the protein/peptide sequence, A and B are water and organic solvent molecules, $X^0{}_i$ is an energy of an interaction between stationary phase and the biomacromolecule, and $N_B(V)$ reflects a change in the mole fraction of organic component of the mobile phase in the course of a gradient elution. Note that the main assumption behind the BioLCCC model is that the interactions between the residues and the surface are short-range "local" interactions. Another assumption of the model is that the interactions between the binary solvent and the surface are limited by Langmuir isotherm approximation resulting in Eq. 1 for the solvent strength ε_{AB}. According to the model, all possible positions of each of the residues in the sequence are described by transition matrices $W(\varepsilon^{eff}{}_i)$, the products of which determine the possible configurations of a biomacromolecule chain of length N (N here is the number of residues) for a particular position of the first residue in the sequence. Respectively, the possible positions of the first residue in the sequence inside the pore are determined by the so-called starting vector P_0 (that is the transition matrix for this residue). The total number of possible configurations for a given sequence will be the sum (using a unit row-vector U^T) of transition matrix products over all possible starting positions of its first residue. This number of configurations is the partition function of a particular sequence inside the pore which, by definition in chromatography, is the distribution coefficient $K_d(V)$ [33]. In case of a gradient separation, the distribution coefficient $K_d(V)$ defines the retention volume. In this equation, V_p and V_0 are the pore and interstitial volumes, respectively. Using the formalism of the BioLCCC model described by the system of equations above one can predict retention of any peptide or protein of known sequence separated under gradient and/or isocratic conditions in the whole range of mobile phase compositions. Note, that at the microscopic level of description the BioLCCC model operates within the assumption on the flexibility of the macromolecule. This assumption considers the macromolecule as a random coil applicable to a certain class of the proteins which are in a denatured state under acidic RP

Table 3 Cytochrome *c* sequences for *equine*, *bovine*, and *canine*

Equine	GDVEKGKKIF	VQKCAQCHTV	EKGGKHKTGP	NLHCLFGRKT	GQAPCF*T*YTD	ANKNKGITW*K*	EETLMEYLEN	PKKYIPGTKM	IFAGIKK*KT*E	R*E*DLIAYLKK	AT*N*E
Bovine	GDVEKGKKIF	VQKCAQCHTV	EKGGKHKTGP	NLHGLFGRKT	GQAPGF*S*YTD	ANKNKGITF*G*	EETLMEYLEN	PKKYIPGTKM	IFAGIKK*KG*E	R*E*DLIAYLKK	AT*N*E
Canine	GDVEKGKKIF	VQKCAQCHTV	EKGGKHKTGP	NLHGLFGRKT	GQAPGF*S*YTD	ANKNKCITW*G*	EETLMEYLEN	PKKYIPGTKM	IFAGIKK*TG*E	R*A*DLIAYLKK	AT*K*E
	10	20	30	40	50	60	70	80	90	100	

"Point mutations" are rendered in italics

HPLC conditions. For example, most of the globular proteins may satisfy this assumption.

Application of the BioLCCC model to predict retention of intact proteins

Table 1 summarizes the experimental conditions used for the separation of intact proteins in the earlier published works [29–31]. All separations were TFA-based using reversed phases C4, C8, and C18 and a variety of columns of different lengths and internal diameters. The protein sequences were extracted from the UniProt database available at http://www.uniprot.org. For each of the protocols and separation conditions, the retention times for extracted sequences were calculated using TFA-based version of the BioLCCC software TheorChromo at http://theorchromo.ru. Both experimental and predicted retention times using different retention models are shown in Table 2. In this table, all experimental retention times obtained for different separation systems and conditions were normalized to the conditions of system #5 described in Table 1. The normalization procedure was introduced earlier and relied on the multiple-point normalization approach [34, 35]. In brief, this approach is based on the assumption that within a predefined set of HPLC conditions the experimental retention times linearly correlate with each other. Note that this assumption is valid in a range of HPLC conditions that have to be identified experimentally. In general, the wide variations in these conditions (column parameters, chemistry of reversed-phase, gradient profile, etc.) may result in nonlinear behavior of the separation selectivity [26, 35, 36]. Therefore, the basis for this normalization is the presence of strong linear correlation between experimental times with the determination coefficient R^2 of 0.98 to 0.99 for the datasets used in this study (not shown). Using the equation $y=ax+b$ with coefficients a and b obtained for the correlation between experimental times one can convert all experimental times to the ones corresponding to the conditions used in system #5. After the conversion, we then calculated the average experimental RTs for each of the

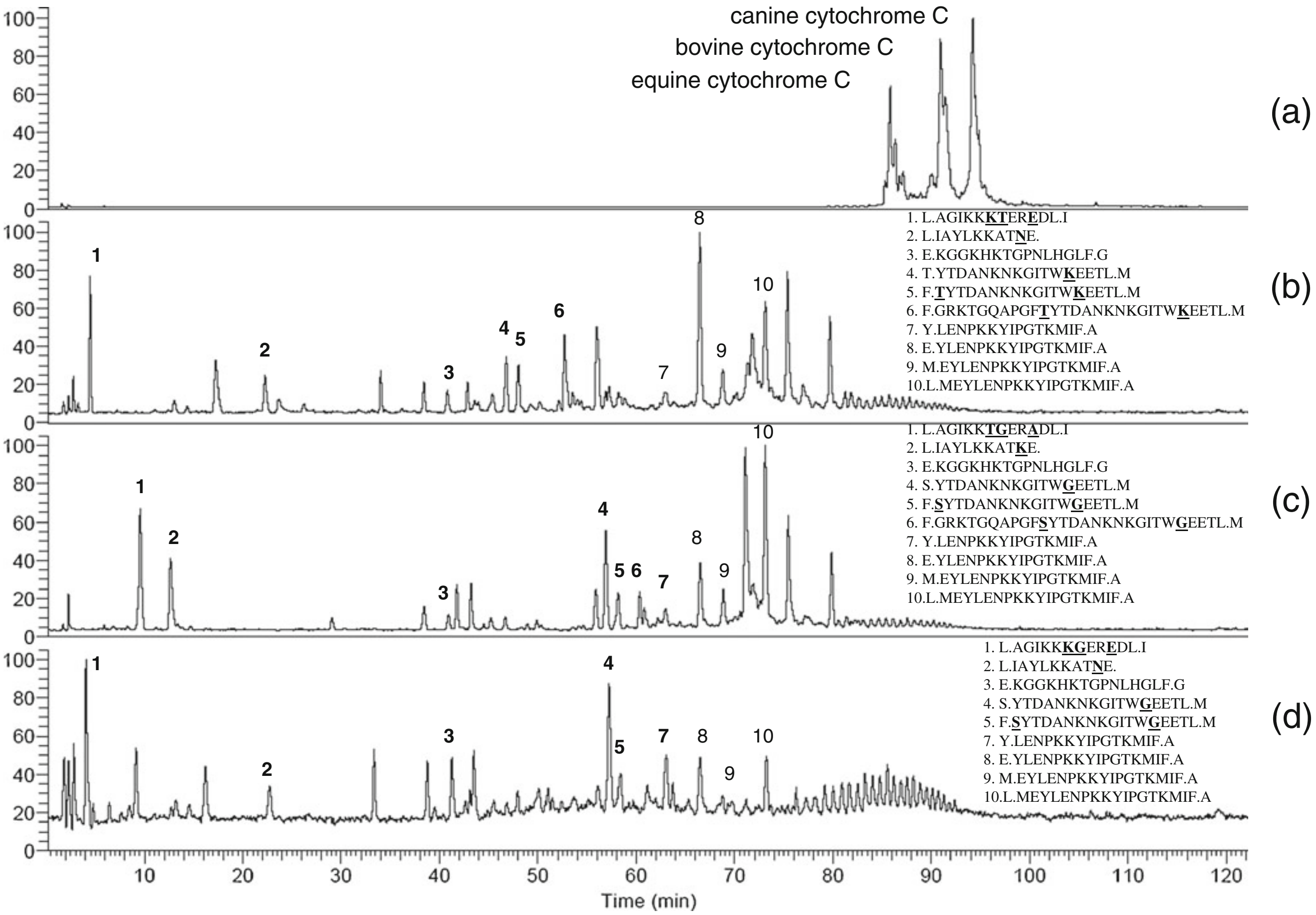

Fig. 2 Chromatograms of HPLC separation for a mixture of cytochrome *c* proteins (**a**) and the proteolytic peptides obtained using pepsin for *equine* cytochrome *c* (**b**); *canine* cytochrome *c* (**c**); and *bovine* cytochrome *c* (**d**). HPLC conditions used in this study were the following: mobile phases A: 5% ACN+0.1% FA and water; B: 95% ACN+0.1% FA and water, linear gradient of 0–35% B in 120 min, flow rate of 0.2 ml/min, ambient temperature; column: Discovery BIO WidePore C18300 Å, 150×2.1 mm, 5 μm particle size

Table 4 Experimental and predicted retention time (min) for cytochrome *c* protein sequence variants

Organism	Retention times ($RT_{exp}/RT_{BioLCCC}$), min		
	Gradient profile		
	0–35%B in 120 min	0–35%B in 56 min	0–35%B in 35 min
Equine	86.29/89.04	42.4/43.29	27.26/28.04
Bovine	91.97/90.84	44.84/44.14	29.02/28.59
Canine	93.36/92.44	46.17/44.89	29.47/29.04
Error (min)	2.21	1.21	0.70

proteins from all available experimental runs. Similar normalization procedure was used for retention times predicted using different retention models. Thus, Table 2 shows both experimental and predicted retention times in the universal scale allowing direct comparison of the absolute RT values. Figure 1 represents comparison of the correlations between experimental and predicted RTs for different models. The BioLCCC model gives the best correlation R^2 of 0.89 (Fig. 1a). Expectedly, the additive model [22] using the summation of retention coefficients (RC) corresponding to each of the residues comprising the sequence, $\sum_i \mathrm{RC}_i$, underperforms the other models as seen in Fig. 1b. It was known long time ago that the RC-based model can correctly predict RTs for relatively short macromolecules not exceeding 10 to 15 monomers (or residues in case of the biomacromolecules) [37]. To correct the effect of sequence length on retention time two approaches have been proposed. In the first approach, the effect of the sequence length was taking into account by using the equation $\mathrm{RT} \sim \sum_i \mathrm{RC}_i \times (1 - m\ln(N))$ [37], where N stands for the number of monomers (residues) in the sequence and m is empirically derived coefficient equal to 0.21 [36]. The applicability of this equation is limited by the sequences not exceeding ~50 residues. In the other approach by Sakamoto et al. the sequence length was taken into account by the equation $\mathrm{RT} \sim \ln\left(1 + \sum_i \mathrm{RC}_i\right)$ [38]. We have found that the latter approach significantly improves the correlation between experimental and predicted RTs for the proteins under study. Figure 1c shows the results of its application for the proteins under study. Note, that in the correlation shown in Fig. 1c we used retention coefficients obtained in [38]. These coefficients are different from the ones obtained by Meek et al. for the original additive model [21, 22] and later by Gilar et al. for the additive model with sequence length correction [37]. Finally, Fig. 1d shows also the result of using TFA-based SSRCalc model developed by Krokhin et al. [23] (on-line version 3.2.3 at http://hs2.proteome.ca/SSRCalc/SSRCalc32.html). Being the most accurate empirical model for prediction of RTs of tryptic peptides it also performs better than any of the additive models when applied to the proteins. Contrary to empirical and additive approaches, the BioLCCC theory relies on the consideration of the large macromolecule as a Gaussian coil. Because the proteins can be better described by the Gaussian coil approximation compared with the short peptides it is not a surprise that protein retention is predicted by this model with the relatively high accuracy. However, one should understand that this approximation is not applicable for the proteins in their native globular form. Obviously, the chromatographic retention times for proteins in their native form will not be sequence-specific and, therefore, the high correlation between the sequence-specific retention time predictions using the BioLCCC model and the experimental retention times indicates the coil-like conformation of the proteins separated in the cited works under acidic conditions of pH~2.0.

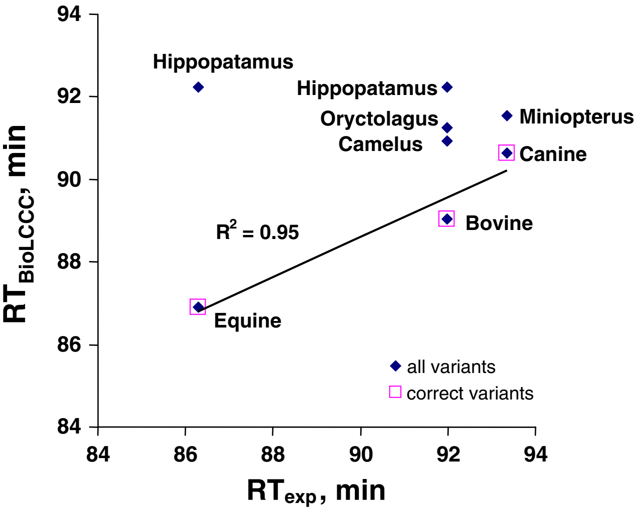

Fig. 3 Correlation between experimental and predicted retention times for cytochrome *c* proteins identified from bottom-up experiments with protein digests obtained using pepsin IMER. Also shown are the protein candidates suggested by Mascot. HPLC conditions used are: mobile phases A: 5% ACN+0.1% FA in H_2O; B: 95%ACN +0.1%FA in H_2O; linear gradient of 0–35%B in 120 min; flow rate of 0.2 ml/min; ambient temperature; RP column Discovery BIO WidePore C18 300Å, 150×2.1 mm, 5 µm particle size

Protein sequence mutations

For the purpose of this study, we have selected proteins with slightly different sequences, referred here as sequences with the "point mutations": cytochrome *c* proteins originated from *equin*e, *bovine*, and *canine*. These proteins are also interesting as they reveal the fundamental sequence specificity of the chromatographic data. Note, that the MS-based identification of the proteins with "point mutations" is especially challenging because the corresponding enzymatic digests are highly cross-correlated and contain large number of identical proteolitic peptides. Cytochrome *c* proteins selected have the same sequence lengths (104 residues) and differ by three to six residues as shown in Table 3. These proteins were separated using FA-based HPLC on reversed-phase C18 column followed by MS measurements. Figure 2a shows the total ion chromatograms (TIC) for all three proteins and Table 4 summarizes the corresponding experimental and predicted retention times. The separations were performed using three different gradient profiles shown in Table 4. Interestingly, we observed significantly better correlation between predicted and experimental retention times when using short gradient, 0–35% B in 35 min, compared with the longer ones. We do not have an explanation for this observation at this time that warrants further studies on the matter. Importantly, the BioLCCC model correctly predicted the elution order of the proteins with minor differences in the sequences.

To further reveal how the prediction of chromatographic times of the intact proteins may assist in the protein identifications we applied the bottom-up approach using tandem mass spectrometry. All three proteins were enzymatically digested using pepsin as described in "Experimental" section above. This digestion was followed by a standard reversed-phase LC-MS/MS analysis. Figure 2b, c, and d show TIC chromatograms for the proteolitic peptides for *equine*, *canine*, and *bovine* cytochrome *c* proteins, respectively. The peptides were subjected to collision-activated dissociation (CAD) MS/MS in the LTQ followed by fragmentation data processing and identification using Mascot as a search engine. This processing yielded up to four protein candidates for each of the digests. The results of identifications are shown in Fig. 3 as a correlation between experimental retention times for the proteins and predicted retention times for all protein candidates including the actual ones. The protein candidates corresponding to the cytochrome *c* proteins under study correlate with their respective predicted retention times, whereas the false positive matches are the obvious outliers. Therefore, the described approach can be a useful filter of false positives in the intact protein identifications including both top-down and bottom-up proteomics methods.

Conclusions

The BioLCCC model has been applied to predict retention times of intact proteins having up to 583 amino acid residues in a sequence. Rather high correlations (determination coefficient R^2 of up to 0.9 that corresponds to the Pearson coefficient R of 0.95) between predicted and experimental retention times for proteins separated using different reversed-phase stationary phases and HPLC protocols have been demonstrated. The results obtained show that the BioLCCC separation model provides a realistic picture of HPLC separation of large biomacromolecules such as proteins and, thus, can be used as a good starting point to interpret separation data in terms of minor differences in sequences of intact proteins due to mutation and/or post-translational modifications. In one possible approach, the utilization of predicted retention times for the intact proteins and their comparison with experimental data allows filtering of incorrect protein assignments in top-down proteomics. Potentially, prediction of retention times can be further used to build up the intact protein retention time database that can be used in protein identification, the approach similar to accurate mass and time tag "shotgun" method for peptide identifications [39, 40].

Acknowledgments This work was supported by Russian Basic Research Foundation (grants 11-04-00515, and 09-08-00633), Swiss National Science Foundation (project 200021-125147/1), and Joint Research Project of Scientific & Technological Cooperation Program Switzerland–Russia (grant agreement 128357 between EPFL and INEPCP RAS). MVG also thanks financial support from the Seventh Framework Program of the European Union FP7-HEALTH-2011, project HiSPRA.

References

1. Abu-Farha M, Elisma F, Zhou H, Tian R, Zhou H, Asmer MS, Figeys D (2009) Anal Chem 81:4585–4599
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al (2001) Science 291:1304–1351
3. Taggart R, Starr C (2006) Biology The unity and diversity of life: mutated genes and their protein products. Thompson Brooks/Cole
4. Dettmer K, Aronov PA, Hammock BD (2007) Mass Spectrom Rev 26:51–78
5. Feng X, Liu X, Luo Q, Liu B-F (2008) Mass Spectrom Rev 27:635–660
6. Davis MT, Stahl DC, Hefta SA, Lee TD (1995) Anal Chem 67:4549–4556
7. Kelleher NL, Lin HY, Valaskovic GA, Aaserud DJ, Fridriksson EK, McLafferty FW (1999) J Am Chem Soc 121:806–812
8. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR (1997) Anal Chem 69:767–776
9. Wolters DA, Washburn MP, Yates JR (2001) Anal Chem 73:5683–5690
10. Palmblad M, Ramstrom M, Markides KE, Hakansson P, Bergquist J (2002) Anal Chem 74:5826–5830

11. Norbeck AD, Monroe ME, Adkins JN, Anderson KK, Daly DS, Smith RD (2005) J Am Soc Mass Spectrom 16:1239–1249
12. Baczek T, Wiczling P, Marszall M, Van der Heyden Y, Kaliszan R (2005) J Prot Res 4:555–563
13. Gilar M, Jaworski A, Olivova P, Gebler JC (2007) Rapid Commun Mass Spectrom 21:2813–2821
14. Klammer AA, Yi X, MacCoss MJ, Noble WS (2007) Anal Chem 79:6111–6118
15. Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O (2007) BMC Bioinformatics. doi:10.1186/1471-2105-8-468
16. Goloborodko AA, Mayerhofer C, Zubarev AR, Tarasova IA, Gorshkov AV, Zubarev RA, Gorshkov MV (2010) Rapid Commun Mass Spectrom 24:454–462
17. Baczek T, Kaliszan R (2009) Proteomics 9:835–847
18. Babushok VI, Zenkevich IG (2010) Chromatographia 72:781–797
19. Petritis K, Kangas LJ, Ferguson PL, Anderson GA, Pasa-Tolic L, Lipton MS, Auberry KJ, Strittmatter EF, Shen Y, Zhao R, Smith RD (2003) Anal Chem 75:1039–1048
20. Sanger F (1952) Adv Protein Chem 7:1–7
21. Meek JL (1980) Proc Natl Acad Sci USA 77:1632–1636
22. Guo DC, Mant CT, Hodges RS (1987) J Chromatogr 386:205–222
23. Krokhin OV, Craig R, Spicer V, Ens W, Standing KG, Beavis RC, Wilkins JA (2004) Mol Cell Proteomics 3:908–919
24. Gorshkov AV, Tarasova IA, Evreinov VV, Savitski MM, Nielsen ML, Zubarev RA, Gorshkov MV (2006) Anal Chem 78:7770–7777
25. Gorshkov AV, Evreinov VV, Tarasova IA, Gorshkov MV (2007) Polymer Sci B 49:93–107
26. Perlova TYu, Goloborodko AA, Margolin Y, Pridatchenko ML, Tarasova IA, Gorshkov AV, Moskovets E, Ivanov A, Gorshkov MV (2010) Proteomics 10:3458–3468
27. DiMarzio EA, Rubin R (1971) J Chem Phys 55:4318–4336
28. Entelis SG, Evreinov VV, Gorshkov AV (1986) Adv Polymer Sci 76:129–175
29. Koyama J, Nomura J, Shiojima Y, Ohtsu Y, Horii I (1992) J Chromatogr A 625:217–222
30. Mant CT, Zhou NE, Hodges RS (1989) J Chromatogr 476:363–375
31. Alltech Catalog 300 Deerfield (1993) ILL Chromatography Alltech Assoc Inc
32. Enzymatic assay of pepsin (EC 3.4.23.1), revised 1996 http://www.sigmaaldrich.com
33. Entelis SG, Evreinov VV, Kuzaev AI (1985) Reactive oligomers. Chemistry, Moscow
34. Tarasova IA, Guryca V, Pridatchenko ML, Gorshkov AV, Kieffer-Jaquinod S, Evreinov VV, Masselon CD, Gorshkov MV (2009) J of Chromatogr B 877:433–440
35. Pridatchenko ML, Tarasova IA, Guryca V, Kononikhin AS, Adams C, Tolmachev DA, Agapov AYu, Evreinov VV, Popov IA, Nikolaev EN, Zubarev RA, Gorshkov AV, Masselon CD, Gorshkov MV (2009) Biochemistry (Mosc) 74:1195–1202
36. Gilar M, Xie HW, Jaworski A (2010) Anal Chem 82:265–275
37. Mant CT, Burke TWL, Black JA, Hodges RS (1988) J Chromatogr 458:193–205
38. Sakamoto Y, Kawakami N, Sasagawa T (1988) J Chromatogr 442:69–79
39. Paša-Tolić L, Lipton MS, Masselon CD, Anderson GA, Shen Y, Tolić N, Smith RD (2002) J Mass Spectrom 37:1185–1198
40. May D, Fitzgibbon M, Liu Y, Holzman T, Eng J, Kemp CJ, Whiteaker J, Paulovich A, McIntosh M (2007) J Proteome Res 6:2685–2694