PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Tandem use of X-ray crystallography and mass spectrometry to obtain ab initio the complete and exact amino acids sequence of HPBP, a human 38-kDa apolipoprotein

Hélène Diemer,[1] Mikael Elias,[2,3] Frédérique Renault,[4] Daniel Rochu,[4] Carlos Contreras-Martel,[5] Christine Schaeffer,[1] Alain Van Dorsselaer,[1]* and Eric Chabriere[2,3]*

[1] Laboratoire de Spectrométrie de Masse Bioorganique, Institut Pluridisciplinaire Hubert Curien, UMR 7178 (CNRS-ULP) ECPM, 25 rue Becquerel F67087-Strasbourg-Cedex 2, France

[2] Laboratoire de Cristallographie et Modélisation des Matériaux Minéraux et Biologiques, CNRS-Université Henri Poincaré, 54506 Vandoeuvre-lès-Nancy, France

[3] Architecture et Fonction des Macromolécules Biologiques, CNRS-Université de la Méditerranée, 13288 Marseille, France

[4] Unité d'Enzymologie, Département de Toxicologie, Centre de Recherches du Service de Santé des Armées, 38702 La Tronche, France

[5] Laboratoire de Cristallogenèse et Cristallographie des Protéines, Institut de Biologie Structurale JP EBEL, 38027 Grenoble, France

ABSTRACT

The Human Phosphate Binding Protein (HPBP) is a serendipitously discovered apolipoprotein from human plasma that binds phosphate. Amino acid sequence relates HPBP to an intriguing protein family that seems ubiquitous in eukaryotes. These proteins, named DING according to the sequence of their four conserved N-terminal residues, are systematically absent from eukaryotic genome databases. As a consequence, HPBP amino acids sequence had to be first assigned from the electronic density map. Then, an original approach combining X-ray crystallography and mass spectrometry provides the complete and a priori exact sequence of the 38-kDa HPBP. This first complete sequence of a eukaryotic DING protein will be helpful to study HPBP and the entire DING protein family.

Proteins 2008; 71:1708–1720.
© 2007 Wiley-Liss, Inc.

Key words: human phosphate binding protein; DING protein; amino acids sequencing; missing gene; X-ray crystallography; mass spectrometry; atherosclerosis.

INTRODUCTION

Eukaryotic genomes are extensively sequenced and an increasing number of them are complete or almost complete. All these data are precious for protein studies. However, it subsists a protein family that is systematically absent from genome databases. These 38-kDa proteins have been named DING, according to the sequence of their four conserved N-terminal residues.[1] Despite their intriguingly systematic absence from genome databases, proteins belonging to this family seem ubiquitous since they have been identified in animals (human, monkey, rat, turkey), plants (*Arabidopsis thaliana*, potato, tobacco), and fungi (*Candida albicans, Ganoderma lucidum*).[1–11] Of course, genes coding for this protein family exists. Although no complete eukaryotic *DING* gene is available in genome databases, few partial DNA sequences coding for this protein family have been cloned or identified in unannoted part of genomes.[12] In addition to this systematic missing, another interesting point in genetics concerns sequence conservation. Indeed, some partial nucleotidic sequences reveal an astonishing conservation between distant species such as potato (a higher plant) and *Leishmania major* (a protozoan) of about 90% identity at the nucleotidic level, over more than 600 bp.[9] This fact can not be explained only by an evolutive constraint on the protein function, because of the redundancy of the genetic code. Topology prediction reveals eukaryotic DING proteins should be all

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/.

Abbreviations: CAI, crystal adhesion inhibitor; ESI, electrospray ionisation; ESTs, expressed sequence tags; GLP, germin-like protein; HDL, high density lipoprotein; HPBP, human phosphate binding protein; HPON1, human paraoxonase; LC, liquid chromatography; MALDI, matrix-assisted laser desorption ionization; MS, mass spectrometry; ORF, open reading frames; SBP, solute binding protein; SSP, synovial stimulatory protein; WHP, woodchuck hepatis virus.

Accession Numbers: The atomic coordinates and structure factors of the corrected HPBP structure have been deposited with the Protein Data Bank, accession code 2v3q.

Grant sponsor: Délégation Générale pour l'Armement; Grant number: CO n°010807/03-10; Grant sponsor: C.N.R.S.

Hélène Diemer and Mikael Elias contributed equally to this work.

*Correspondence to: Alain Van Dorsselaer, Laboratoire de Spectrométrie de Masse Bioorganique, Institut Pluridisciplinaire Hubert Curien, UMR 7178 (CNRS-ULP) ECPM, 25 rue Becquerel F67087-Strasbourg-Cedex 2, France. E-mail: vandors@chimie.u-strasbg.fr or Eric Chabriere, Architecture et Fonction des Macromolécules Biologiques, CNRS-Université de la Méditerranée, 13288 Marseille, France. E-mail: eric.chabriere@afmb.univ-mrs.fr.

Received 1 June 2007; Accepted 10 October 2007

Published online 12 December 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21866

© 2007 WILEY-LISS, INC.

similar to phosphate Solute Binding Protein (SBP) associated with bacterial ABC transporter. SBP in eukaryotes have never been predicted or characterized, despite eukaryotic ABC transporters exist. This prediction was recently comforted, since the structure of the first DING protein, the Human Phosphate Binding Protein (HPBP) was solved.[9] This work confirms that DING proteins should be all capable of binding inorganic phosphate with very high affinity and specificity. Moreover, some eukaryotic DING proteins possess other biological activities, which allowed them to be isolated. For example, a DING protein was identified in tobacco by virtue of its strong binding to an *Arabidopsis thaliana* germin-like protein (GLP) named AtGER3.[1] Similarly, Mehta *et al.*[13] isolated a Woodchuck Hepatis Virus (WHV) binding protein from woodchuck plasma which is related to the DING protein family. Unfortunately, most of eukaryotic DING proteins, missing from nucleotidic databases, are only N-terminal sequenced and poorly characterized.

In human, different works report DING proteins identification. There are the Genestein-binding protein from human carcinoma cells,[3] the Hirudin-binding protein from human fibroblasts,[6] the human synovial stimulatory protein (SSP) from synovial liquid,[14] the Crystal Adhesion Inhibitor (CAI) from human kidney cells,[5] and the Human Phosphate Binding Protein (HPBP)[15,16] from human plasma. The Genestein-binding protein and the Hirudin-binding protein are not enough characterized and sequenced to conclude whether their biological activities are due to different human DING protein or not. The three other identified proteins are more sequenced. From sequences alignment, it can be concluded that there is at least these three different DING proteins missing from human genome database. Furthermore, these three proteins, which are more characterized, are related to pathologies. The SSP, isolated from human synovial liquid, possesses autoantigen activity, lymphocyte stimulatory activity, and a putative role in the etiology of rheumatoid arthritis.[14,17] Concerning the CAI, this protein was first identified from green monkey cells, then in human kidney cells and purified from human urine. The protein was two-third sequenced (261 amino acids) by internal peptide digestion and N-termini sequencing. CAI is supposed to prevent from the growth of kidney stone, inhibiting oxalate calcium crystals formation, and probably calcium phosphate nuclei genesis.[5] HPBP is a human apolipoprotein that binds phosphate. HPBP was serendipitously discovered while processing on structural studies on human paraoxonase (HPON1), a 38–45 kDa glycosylated HDL-associated lipoprotein. The structure of HPBP consists in the first DING protein structure available, and provides the first complete eukaryotic DING protein sequence. Crystals of HPBP were obtained from a supposedly pure HPON1 solution. Facts that HPON1 is glycosylated, possesses the same molecular weight as HPBP and that finally these two hydrophobic proteins are strongly associated explain why the copurification of HPBP was not visible on SDS-PAGE analysis.[18] Furthermore, recent studies show that the association of these two proteins is believed to be physiologically relevant as their oligomerization is modulated by calcium and phosphate concentration[18] (Elias *et al.*, in preparation). Even if HPBP physiological involvements are not yet clearly established, concomitant facts suggest that HPBP could be tentatively regarded as a new predictor or as a possible therapeutic agent for phosphate-related disorders, including atherosclerosis.[9,19]

As HPBP is missing from the human genome database, its sequence had to be completely assigned from the electronic density map. This first sequence was helpful for the alignment of peptide sequences subsequently obtained by classical techniques such as N-termini sequencing, internal peptide digestion, and mass spectrometry analysis. The crystallographically predicted sequence was thus confirmed at 75%.[9] However, this sequence obviously still contains some ambiguities. Of course, it is necessary to obtain the complete sequence without any ambiguities to perform biochemical studies, mutational studies, gene search experiments, and de novo *HPBP* gene synthesis.

At the post genomic era, gene and protein sequences are mainly obtained from databases or in few cases from RT-PCR experiments. In the case of HPBP, all these classical approaches failed. Attempting to sequence completely a 38-kDa protein represents a real technical challenge. In fact, completely sequenced protein hardly exceeds 100 amino acids. Chao et al.[20] have completely sequenced the 71 amino acids of Applagin (*Agkistrodon piscivorus piscivorus* platelet aggregation inhibitor) by Edman degradation. Another example consists in the complete sequencing of the 90 amino acids of cytochrome b_{558} from *Ectothiorhodospira vacuolata* by a combination of automated Edman degradation and mass spectrometry.[21] Hellman et al.[22] tried to sequence extensively a trypsin-generated fragment from human complement factor C3 (302 amino acids). They used CNBr cleavage and analyzed the obtained peptides by Edman degradation. They, thus, obtained about 90% of the 302 amino acids of this C3 fragment. For a large protein, only one technique seems not to be sufficient for obtaining the entire amino acids sequence.

Thus, we develop a strategy to successfully sequence the 38-kDa HPBP. A series of enzymatic digestions will be performed on HPBP to generate peptides allowing obtaining a maximum of sequence information by MS fragmentation in LC-MS/MS, MALDI-MS/MS experiments. Data obtained from each digestion experiments will be used in a sequential way to introduce amino acid sequence corrections. The template sequence used first will be the sequence deduced from interpretation of X-ray data. This sequence, containing zone with ambiguous amino acids, will be corrected, step by step using the information from each digestion. Data from digestion with trypsin, Lys-C, chymotrypsin, and thermolysin will be successively used. Finally, the experimental molecular mass measured on HPBP by ESI-MS with an accuracy of ± 0.5 Da will be

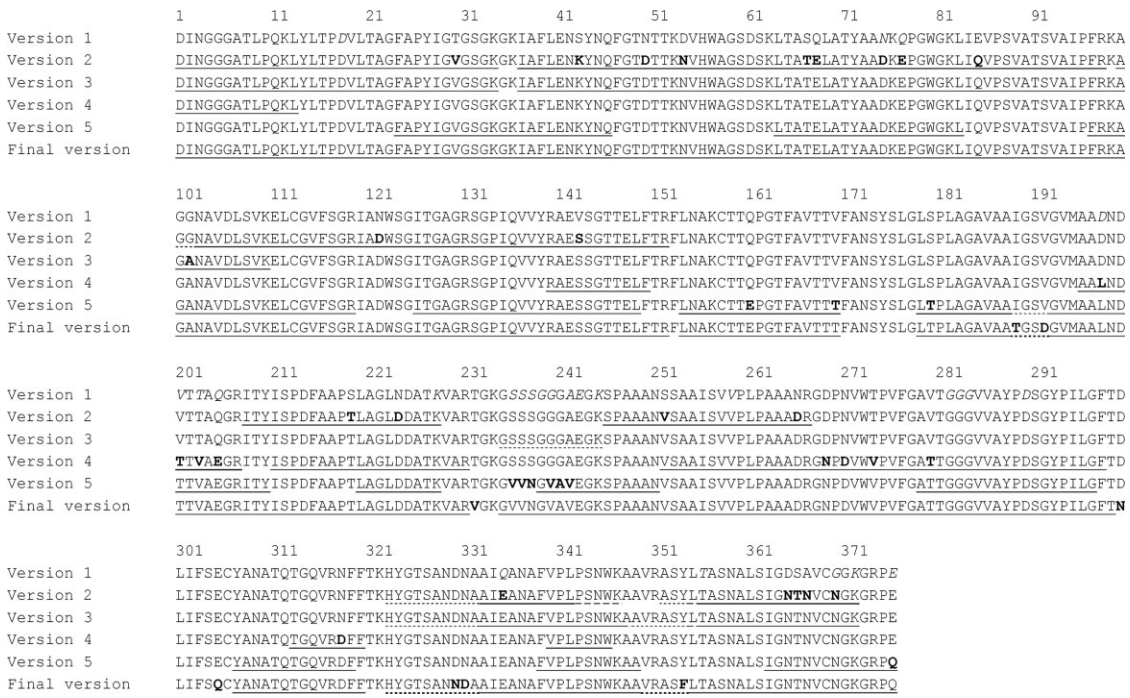

Figure 1

Amino acid sequences of HPBP. Consecutive corrections brought from the first proposed sequence to the final sequence of HPBP. In sequence version 1, Amino acids annotated in italic could be ambiguous. Amino acids in bold correspond to a correction. Underlined amino acids in continuous line are confirmed or corrected and amino acids with dotted line have been seen but not verified.

compared with the mass calculated from the proposed sequence, and confronted to electronic density maps.

This study reports the method used to obtain the complete and unambiguous sequence of HPBP, a 38-kDa protein.

RESULTS

Crystallographic sequencing

Amino acid sequence (sequence version 1 on Fig. 1) has been attributed from the shape formed by the electronic density map [Fig. 2(A)]. Then, the first 15 amino acid have been sequenced by Edman degradation, and the crystallographically predicted sequence was checked at 75% by mass spectrometry, showing an error rate of about 11% for this verified sequence portion.[9] Thus, this sequence probably still contains remaining errors that could be explained, as some ambiguities are possible at the resolution of 1.9 Å. In fact, the electronic density map is related to the electron number of the atoms. And at this resolution, it is not possible to clearly discriminate carbon, nitrogen, and oxygen atoms, as they possess roughly the same number of electrons (six, seven, eight electrons, respectively). This limitation implies that some amino acids possess similar electronic density shapes, such as Asn and Asp, Gln and Glu, and Val and Thr [Fig. 2(B,C)], and are thus difficult to discriminate. Furthermore, some protein residues possess multiple conformations. Such agitation modifies the electronic density shape, and can be a source of confusion. To illustrate, some double serine conformation causes similar electronic density shapes as threonine or valine residues [Fig. 2(C)]. Another cause of ambiguity concerns disordered atoms. In crystallography, disordered atoms contribute less than ordered atoms for diffraction. Consequently, these agitated atoms disappear from the electronic density map. This fact mainly concerns residues located at the protein extremities or surface, and can be attributed to some thermal or statistic disorder.[23] This fact causes truncated electronic density, which can be assimilated to the density corresponding to shorter residue [Fig. 2(D)].

Although the electronic density map possesses intrinsic limitations, other criterion can be used to increase the discrimination level. The first concerns chemical environment analysis. As previously said, confusion is possible between Val, Thr, or Ser residues. Since these amino acids do not possess same chemical properties, environment analysis can allow their discrimination. In fact, valine residue is more likely to fit in hydrophobic neighborhood, whereas threonine or serine residue can accommodate more polar contacts such as water molecules or

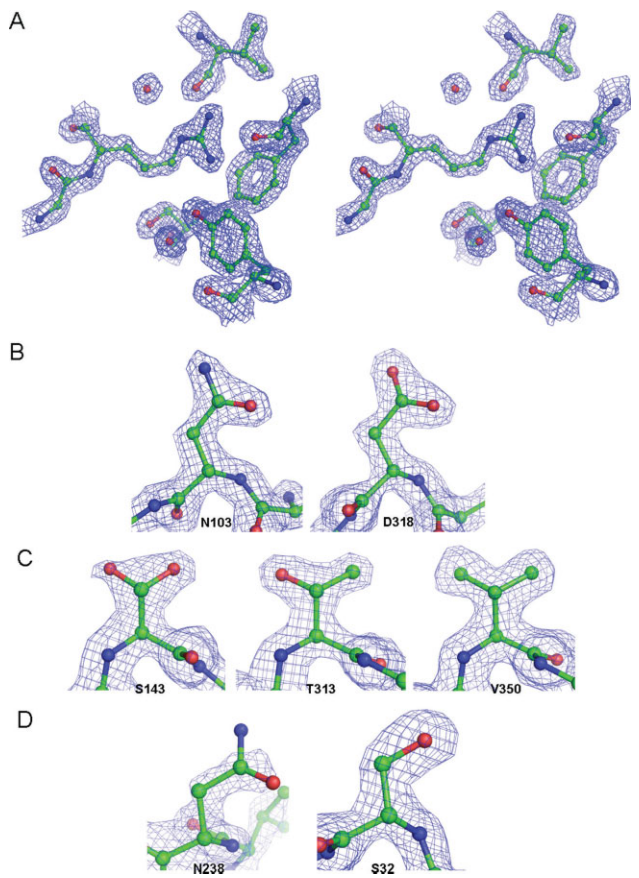

Figure 2

*Electronic density map at 1.9-Å resolution: possible ambiguities. (**A**) Stereo view of a ball-and-stick representation of R374 region in HPBP structure at 1.9-Å resolution. The electronic density map $2f_o - f_c$ is contoured at 1.75σ. (**B**) Comparison between electronic density shapes of N103 and D318. This figure shows that Asn and Asp possess similar shape at 1.9-Å resolution. The electronic density map $2f_o - f_c$ is contoured at 1.5σ. (**C**) Comparison between electronic density shapes of V350, T313, and S143. This figure shows Val and Thr, but also the serine double conformation, possess similar shape at 1.9-Å resolution. The electronic density map $2f_o - f_c$ is contoured at 1.5σ. (**D**) Comparison between electronic density shapes of N238 and S32. N238 residue is not well-defined in the electronic density map, possibly caused by an agitation along the C_β–C_γ axis. This truncated density can be misattributed to a serine residue. The electronic density map $2f_o - f_c$ is contoured at 1.5σ. All residues are represented in ball-and-sticks, with carbon atoms in green, oxygen atoms in red, and nitrogen atoms in blue.*

protein polar atoms [Fig. 3(A)]. Another criterion can be provided by checking the refined crystallographic temperature factor (B factor). B factor is the quadratic standard deviation of the atomic thermal motion,[23,24] and is correlated, at this resolution, to the occupancy and the electron number. As a consequence, if an atom is badly placed, refined B factor will tend to compensate, increasing or decreasing. To illustrate, in the case of Asn/Asp confusion, an oxygen atom (eight electrons) is put at a wrong place, instead of a nitrogen atom (seven electrons). Refinement programs tend to compensate the excess of electron (+1) by increasing the thermal motion, the B factor. As these residues possess rigid chemical function (like Glu and Gln, for example), B factor values should be of the same order for equivalent motioned atoms. A large variation of B values suggests an attribution error, like an oxygen atom in place of a nitrogen atom or vice versa. So, it is possible, in some cases, to discriminate Asp and Asn, and Gln and Glu. All errors are recapitulated in Table I.

Thus, the use of these additional criterions can improve the crystallographic sequencing, but a huge work is required. We deduced from these facts that the more judicious strategy to attribute the complete sequence of HPBP with high confidence was to perform mass spectrometry experiments, and to combine both techniques.

Mass spectrometry sequencing

Mass measurement of intact HPBP

A mass measurement was first performed on intact solubilized HPBP. The molecular weight of the intact protein with its two disulfide bridges (Fig. 4) was mea-

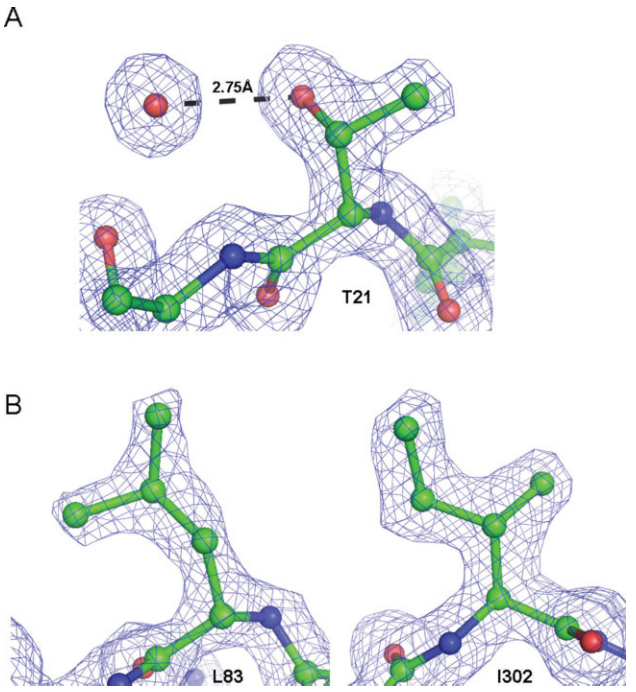

Figure 3

*Discrimination of amino acid type analyzing electronic density map. (**A**) Ball-and-sticks representation of T21. Although Thr possesses similar electronic density shape as Val at 1.9-Å resolution, T21 was successfully discriminate by examination of its environment. In fact, one residue atom is closed to a water molecule, suggesting this residue atom is polar and hydrogen bonded to the water molecule. The distance is indicated in Angströms (Å). The electronic density map $2f_o - f_c$ is contoured at 1.5σ. (**B**) Ball-and-sticks representation of L83 and I302. Although Leu and Ile are rather hard to discriminate by mass spectrometry, this figure clearly shows these two residues possess very different electronic density shapes. The electronic density map $2f_o - f_c$ is contoured at 1.5σ. All residues are represented in ball-and-sticks, with carbon atoms in green, oxygen atoms in red, and nitrogen atoms in blue.*

Table I
Crystallographic Sequencing Errors

Residue type (occurrence in the final HPBP sequence)	D(17)	N(21)	V(33)	T(36)	E(10)	Q(8)	A(55)	S(25)	L(24)	K(17)	F(17)
Confounded with (number of errors compared to the final HPBP sequence)	N(8), S(1), V(1)	D(6), S(1), A(1), G(1)	T(4), S(3), G(1), A(1)	S(4), V(3), I(1)	Q(5)	E(4)	G(2)	V(1)	D(1)	S(1)	Y(1)
% of attribution confidence	41.2	57.1	72.7	77.8	50	50	96.4	96	95.8	94.1	94.1

Type of amino acid is indicated by its one letter code. G, I, P, W, R, M, C, and H are not indicated in the table, as they have been perfectly attributed from crystallographic data. % of attribution confidence is calculated for each residue as (number of correct attribution/residue occurrence) × 100.

sured at 38529.11 ± 0.06 Da. The ESI spectrum is shown on Figure 5. The final proposed sequence for HPBP will have to fit with this experimental value.

Determination of the correct sequence of HPBP by parallel enzyme digestions

General strategy

The general strategy used is summarized in Figure 6. The first crystallographic sequence (sequence version 1 in Fig. 1) acted like an "Ariane wire" for alignment of peptide subsequently obtained by different digestions of HPBP and analyzed by mass spectrometry.

The HPBP was obtained from 1D gel bands, to insure to have a protein as pure as possible. Excised bands were digested with the following enzymes in parallel: trypsin, Lys-C, chymotrypsin, and thermolysin. The resulting digests were all analyzed by two mass spectrometry techniques: MALDI-MS/MS, and nanoLC-MS/MS. Interpretation of the obtained MS/MS data was used in the following way : the sequence proposed from X-ray crystallography (sequence version 1) was successively modified so that it fits with sequence information obtained from each of the digestion experiments, in a complementary way. Finally, the molecular mass calculated from the last sequence version obtained was compared with the experimental molecular mass. For clarity, modifications made on each sequence version will be shown in bold in Figure 1. For each digestion, peptide sequences obtained from MS/MS fragmentations will be reported on a specific table (supplementary data Tables III–VI, respectively for trypsin, Lys-C, chymotrypsin, and thermolysin digestions).

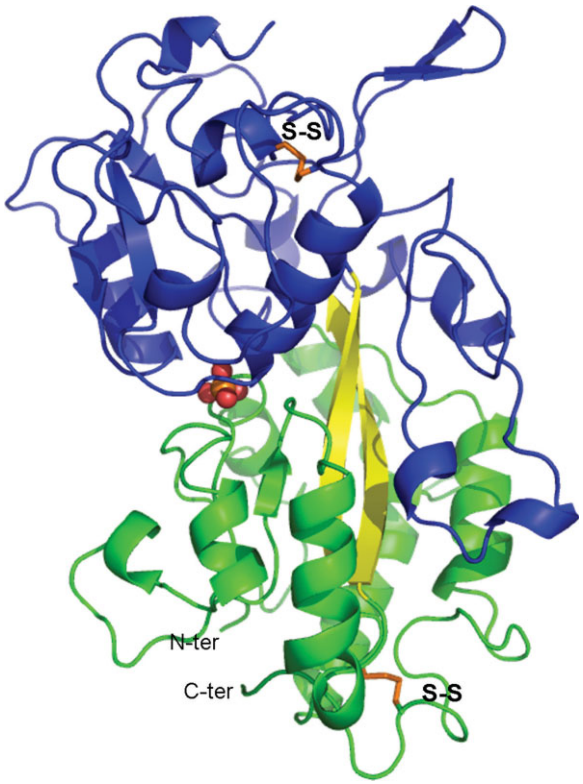

Figure 4
Corrected X-ray structure of HPBP. HPBP possess an elongated fold composed of two adjacent globular domains (in blue and green). Each domain is constituted by a central β-sheet core flanked by α-helices and contains a disulfide bridge (C113-C158 and C306-C359, orange sticks). Interconnected by an antiparallel two-stranded β-sheet acting as a hinge (in yellow), the two domains form a deep cleft wherein is bound a phosphate molecule (red balls).

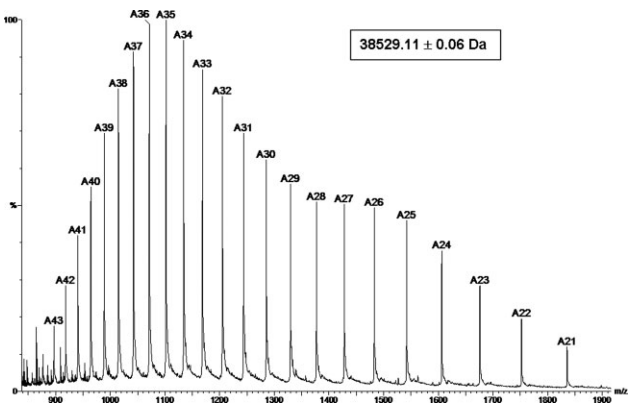

Figure 5
ESI spectrum of intact HPBP, acquired on a Q-T of two mass spectrometer. The molecular mass was calculated from the m/z values obtained from the series of observed multiply charged ions and yielded an average value of 38 529.1 Da with a standard deviation of ±0.1 Da.

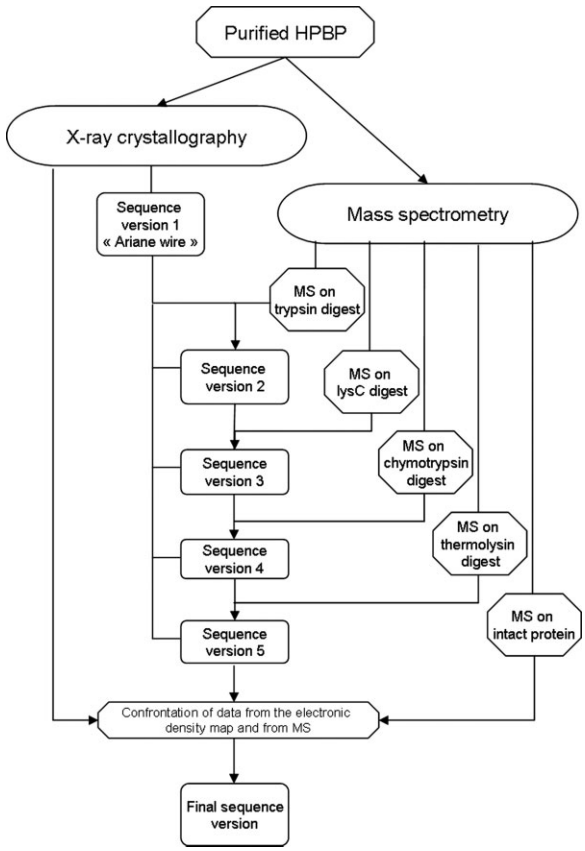

Figure 6

General strategy. Schematic view describing the sequencing of HPBP by mass spectrometry, using the X- ray crystallography data as an "ariane wire". Different versions of the HPBP amino acid sequence are presented and the modifications from version n to version n + 1 correspond to corrections brought by MS/MS data. The last version of the sequence is characterized by a calculated molecular mass corresponding to the experimental mass as measured by ESI-MS.

In-gel tryptic digestion. MALDI-MS/MS on the total digest yielded sequence information for nine peptides (T13, T11, T5, T4(2), T12, T14, T8, T6-7, and T21) and nanoLC-MS/MS on 17 peptides (T14(1), T13, T11, T10, T5, T24(2), T1, T12, T14, T6, T8, T6-7, T21, T26, T17, T2, and T24). These data allowed to bring 20 corrections on sequence version 1 (obtained from X-ray crystallography) (see supplementary data Table III). The corrected sequence is presented in Figure 1 and is named sequence version 2.

The MALDI-MS spectrum of the tryptic digestion peptides of HPBP is shown in Figure 7. The peptides matching with sequence version 1 of HPBP are marked with a star.

In-gel Lys-C digestion. The nanoLC-MS/MS analysis allowed to obtain sequence information on 11 peptides. For seven of them (L4, L6, L5, L7, L9, L7-8, and L2), these data confirmed the sequence version 2 (supplementary data Table IV). For three peptides, these data allowed to introduce corrections (L10, L17, and L16) or to identify peptides. The last peptide (1027.56 Da) was not identified at this point. The corrected sequence is presented in Figure 1 and is named sequence version 3.

In-gel chymotryptic digestion. The MALDI-MS/MS and nanoLC-MS/MS analysis allowed to obtain sequence information on 29 peptides (supplementary data Table V). For 21 of them, these data confirmed the sequence version 3. For eight peptides, these data allowed to introduce corrections (start-end: 313–320, 268–277, 196–207, 267–277, 266–277, 264–277, 278–298, and 258–277). Finally, nine corrections have been brought to obtain the version 4 of HPBP sequence, shown in Figure 1.

In-gel thermolysin digestion. Both MS/MS techniques allowed to obtain sequence information on 31 peptides. For 26 of them, these data confirmed the sequence version 4 (supplementary data Table VI). For five peptides, these data allowed to introduce corrections (start-end: 240–251, 154–166, 363–376, 179–197, and 154–170). Ten amino acids have been corrected to obtain the sequence version 5 of HPBP (see Fig. 1).

To obtain more peptides, in a separate experiment, an in-gel consecutive digestion with trypsin and then thermolysin was performed. Overnight digestion with trypsin, followed by a 2-h digestion with thermolysin allowed to clearly identify and sequence peptide: 307–317 (YANATQTGQVR) where the first three amino acids were unambiguously sequenced for the first time by mass spectrometry, confirming the data proposed by X-ray crystallography. The other eight amino acids were confirmed.

Finally, 95% of amino acids were sequenced by MS/MS and 95% of them (or 90% of the amino acid sequence) have been confirmed or corrected.

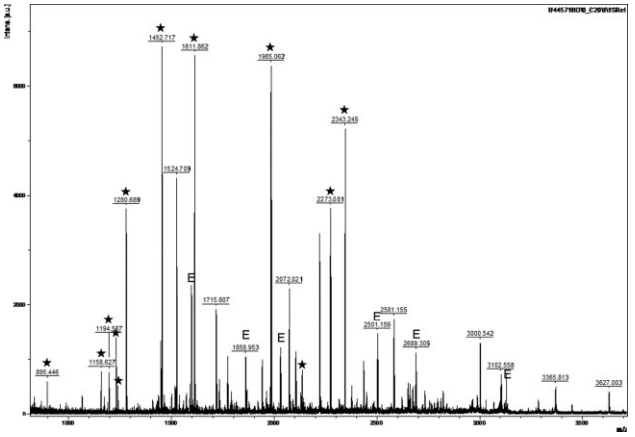

Figure 7

MALDI-MS spectrum of trypsin digest of HPBP. The peptides matching with sequence version 1 of HPBP are marked by a star.

Table II
Confrontation of Data from Electronic Density Map and from Mass Spectrometry

Start-end	Sequence version 5	Mass difference (Da)	Unambiguous amino acids deduced from X-ray data	Amino acids corrected	Final sequence
189–192	IGSV	4	190G 191S	V192D I189T	TGSD
352–354	ASY	−16	352A 353S__Y354F	Y354F	ASF
323–332	HYGTSANDNA	0	323H 324Y 325G 327S 328A	D330N N331D	HYGTSANNDA
153	F	−4	153F	—	F
171–178	FANSYSLG		171F 172A 174S 175Y 176S 177L 178G	—	FANSYSLG
232–234	TGK		233G 234K	T232V	VGK
299–306	TDLIFSEC		299T 301L 302I 303F 304S 306C	D300N E305Q	TNLIFSQC
321–322	TK		322K	—	TK

Combination of the two techniques

All these corrections lead to a complete HPBP sequence (version 5) whose calculated molecular weight (38544.9 Da) is different from the experimental value obtained by ESI-MS (38529.1 Da) (Fig. 5). This is explained by the residual 10% amino acids, which still contain ambiguities from crystallographic sequencing. Half of them have been seen in mass spectrometry data (portions 189–192, 352–354, and 323–332) and the other half was not sequenced, because it corresponds essentially to hydrophobic peptides (portions 153, 171–178, 232–234, 299–306, and 321–322). But fortunately, mass differences have been determined on sequenced portions from MS/MS data. These mass values associated to the theoretical and experimental mass comparisons allowed to deduce the mass difference on the MS unknown portions. These data were then confronted with X-ray data (electronic density map, chemical environment analysis, and B factors). All these constraints issued from both techniques allowed to establish correct sequences (Table II). For example, a difference of 4 Da between experimental and theoretical masses has been deduced on the portion Ile-Gly-Ser-Val (189–192) from the MS/MS data obtained for the peptide 179–197 generated by the thermolysin digestion (see supplementary data Table V). According to the MS/MS spectrum, a Thr is present in position 189. This fact has been confirmed by X-ray data. So, the mass difference on the portion changes to 16 Da. Since amino acids Gly190 and Ser191 are nonambiguous from X-ray data, the only possibility to match the mass of 16 Da, is to have an Asp in the place of Val192. Portion 189–192 can therefore only be Thr-Gly-Ser-Asp.

For the major part of HPBP, the sequence was established using both X-ray and MS data which were converging, as represented in Figure 6. The complete protein was sequenced from the electronic density map, and about 90% amino acids were clearly verified using mass spectrometry. The crystallographic sequence and the MS/MS data of all peptides issued from five different enzymatic digestions provide enough information to obtain the exact sequence of HPBP. Using this sequence, we correct the three-dimensional model of HPBP, and this accurate model fits clearly the refined electronic density map. Furthermore, this final sequence perfectly converges to the experimentally measured mass with a difference in mass of only 0.2 Da (38528.9 and 38529.1 Da, respectively).

DISCUSSION

Crystallographic sequencing relevance

Comparison between the first crystallographically obtained sequence and the final sequence shows that some residues (8/20) have been perfectly attributed: Gly, Ile, Pro, Trp, Arg, Met, Cys, and His. Most of these residues possess characteristic shape. For example, Trp present a unique shape among amino acids and can be attributed with very high confidence, regarding X-ray data. Phe also possesses unique shape, and there is normally no ambiguity to attribute this residue. However, Phe was confounded one time with Tyr, and this clearly corresponds to a human mistake. Other errors can be observed concerning amino acids that possess a priori unique shape, such as Lys and Ala. These errors are due to disordered side chains that disappear from electronic density map. As agitation causes truncated electronic density, these residues have been confounded with shorter residues (Lys with Ser, and Ala with Gly). However, these errors are rare, and Ala and Lys have been successfully attributed in most cases (over 90%). Some other amino acids possess roughly the same shape, such as Val, Thr, or double conformation of Ser [Fig. 2(C)]. As previously said (See "Results" section), Val is rather a hydrophobic residue whereas Thr and Ser can accommodate a more polar environment. This criterion, based on chemical properties of neighbor atoms, seems to be reliable, since Val and Thr were correctly attributed in over

70% cases. As multiple conformation does not concern majority of protein residues, Ser was confounded with Val only one time, and was successfully attributed in most cases (96%). A similar problem can be observed with Leu. In fact, Leu residue possesses resembling electronic density shape with Asn or Asp. Similar criterion based on hydropathy of neighborhood was used, and allowed to discriminate Leu with efficiency (over 95%). The more problematic residues are Asp, Asn, Glu, and Gln. Apart from the cases where side chains were disordered, they were always confounded with their isosteric counterpart (Asp with Asn; Glu with Gln). Indeed, they possess also roughly similar shape and can hardly be discriminated regarding neighborhood chemical properties. In fact, they are all capable to fit hydrogen bonds network. Thus, they can not be attributed with confidence at 1.9-Å resolution (around 50% of correct attribution). At the beginning of the crystallographic sequencing, we do not use B factor value as criterion to discriminate nitrogen atoms from oxygen atoms (See "Results" section). Indeed, this criterion requires hard work on three-dimensional model. This suggested that the more judicious strategy to improve sequence quality was to perform mass spectrometry experiments. In fact, B factor value was usefully checked to confirm mass spectrometry data in the last step. As this criterion seems really reliable confronting with mass spectrometry data, it can be noted that the accuracy of the crystallographically predicted sequence might have been much higher. Finally, the exact HPBP sequence reveals that the error rate of the first crystallographic sequencing is about 14% (3% more than an intermediate sequence[9]). This result shows that numerous ambiguities are possible using crystallographic sequencing, but also shows this sequence provides lot of informations, and is really reliable for alignment of peptides subsequently obtained by MS.

Mass spectrometry sequencing

Since the wide use of the genomic sequences to identify proteins using the proteomic approach, there are very few examples of direct protein sequencing by mass spectrometry. The almost unique approach used today to obtain the sequence of a protein, consists in using the protein databases to search for a sequence that could yield similar MS/MS for a few digestion peptides. The pitfall of this approach is the fact that a protein sequence could have escaped a proper genome translation. The scientific community is now aware that protein databases, produced by automatic translation of genome sequences could contain errors,[25] which could correspond to the following cases: (i) start codon are in wrong position, or are not detected,[26] (ii) ORF are not predicted in the case of very small genes,[27] (iii) Interrupted CoDing Sequences in prokaryotic genomes (ICDD) yield frame shifts.[28,29] In addition to these clearly established annotation or sequencing errors problems, some sequences are missing from certain genomes, as we discuss later.

For all these reasons, it is important to continue to use and develop techniques for direct protein sequencing. The state of the art methodology in mass spectrometry allows to make possible the sequencing of a protein up to 40 kDa using alignment information from other sources (X-ray crystal structure, homologies). Such a study would require a few picomoles of a purified 40-kDa protein. Digestions can be performed on the protein in solution if it is soluble without detergents incompatible with proteolytic enzyme activity. Otherwise, the protein can be separated on a 1D band or a 2D gel spot, and the digestion made "in gel". In the case of HPBP, the 1D gel step was only used on the already purified protein to remove all detergents and salts. Also the "in-gel" digestion can be performed on very small amounts. Digestions using enzymes with different specificity would yield peptides with different properties (solubilization, size, polarity) increasing the chances of sequencing all amino acids. The mass spectrometry instrumentation should be the same as in classical proteomic analysis in term of sensitivity, MS/MS capability, resolution and mass accuracy, automation, ease of use. It is also necessary to widely use the complementarities between sequence information given by MALDI-MS/MS and LC-ESI-MS/MS as illustrated in this work. Also, the de novo automatic sequence of MS/MS spectra must be used systematically, because there are still ambiguities and/or errors in the putative sequences. Therefore, such a protein sequence determination requires all the up to date tools of proteomic analysis. Edman degradation has still its utility since its sensitivity can be in the range of 1 picomole and it can yield an N-terminus sequence of 10–20 amino acids. The technique is obviously limited by digestion problems in presence of detergents and the need for isolating 100% pure digestion peptides. Unfortunately, Edman degradation instruments are now seldom and the expertise is lost in many laboratories. On the contrary, the use of mass spectrometry is constantly in progress.

Finally, this work shows that the direct sequence determination of proteins is useful and possible using the classical and wide-spread proteomic analysis methodology based on mass spectrometry.

Mass spectrometry and X-ray crystallography are complementary

The determination of a 376 amino acids sequence by direct methods is rather unusual nowadays but is still very challenging. Such an approach was routine for smaller proteins in the 80s and the 90s but are today rarefied by the large amount of available nucleotidic sequences. This ab initio sequencing clearly reveals that only one technique does not bring enough information

to sequence completely in an unambiguous way a large protein.

The primary sequence obtained by X-ray crystallography was used like an "Ariane wire", useful to align peptide sequences subsequently obtained by mass spectrometry, without the need for having overlapping peptides. It can be noted that X-ray crystallography technique gives important information that can hardly be obtained using mass spectrometry, such as the exact number of amino acids and the presence of the disulfide bridges. Mass spectrometry experiments were used to correct errors from crystallographic sequencing (residues similar in shape, agitated, or in multiple conformations, see "Results" section), and about 90% of the primary sequence was thus verified. On the other hand, classical ambiguities in mass spectrometry sequencing were solved using X-ray data, as crystallography allows to discriminate amino acids similar in mass (i.e., Ile and Leu, Lys, and Gln), but different in shapes [Fig. 3(B)]. Finally, HPBP was doubly sequenced: the protein was first completely sequenced using X-ray crystallography, and then, almost 90% of amino acids were verified by mass spectrometry. The last 10% were rechecked using X-ray data and mass information. This strategy allowed us to determine the complete sequence of the 38-kDa HPBP (final version) that clearly fits with all data produced. This work represents, to our knowledge, the first ab initio complete sequencing of a protein of this size.

Genomic aspect

The complete and unambiguous HPBP sequence we obtained confirms *HPBP* gene is neither in the sequenced human genome, nor in any genomic databases. Although this is rather a rare case, this is not the only sequence missing from genomic databases. For example, only concerning DING proteins, at least three different proteins are missing from the sequenced human genome.[9] There is HPBP, but also SSP, isolated from human synovial liquid,[14] and CAI, identified in human kidney cells.[5] Maybe more DING proteins are missing, like the Genestein-binding protein from human carcinoma cells[3] and the Hirudin-binding protein from human fibroblasts,[6] but they are not enough sequenced to conclude if their biological activities are due to different human DING proteins or not.

There are some possibilities to explain the absence of a gene sequence in a genome database. Sometimes the gene refers to "noncoding" genome sequence. In fact, prediction of Open Reading Frames (ORF) position and number can be difficult, especially concerning eukaryotic genomes. Nevertheless, when the protein is found, for example using proteomic techniques,[30] or studying Expressed Sequence Tags (ESTs),[31] the corresponding gene can be found in "noncodant" nucleotidic databases. Indeed, the protein or mRNA sequence allow to find the badly annoted gene. In other cases, the gene refers to a sequence absent from the database. This absence from supposedly complete genome database raises one question. Are the completely sequenced genomes really complete? One possible explanation to such an absence concerns heterochromatin and high repeated regions. In fact, these genome regions are difficult to sequence using classical techniques, that is, the human genome is generally said to lack heterochromatin sequence.[32] The missing genes can be located in these poorly sequenced loci. Some studies, which reports new genes discovery support this hypothesis. In fact, some completely unknown genes discovered in *Arabidopsis thaliana* present similarities with transposons and/or have features of repeated sequences.[33] These facts confirm "complete" genomes still contain significant gaps.

As previously said, all three human DING proteins are missing from the sequenced human genome. Only a small human EST (60 amino acids; Berna *et al.*, in preparation) is available for these proteins, other available human sequences are peptidic sequences. In addition, all other eukaryotic DING genes are missing from nucleotidic database too, whether it is from rat, potato, or *Leishmania major*. This systematic absence might not be a coincidence. However, some partial genes coding eukaryotic DING proteins from different organisms are available in databases or recently partially cloned. They apparently do not show splicing tendency, as all identified partial genes are contiguous (Berna *et al.*, in preparation). The systematic absence of complete eukaryotic DING genes in databases, added to the astonishing nucleotidic conservation observed between distant species[9] possibly suggests a more fundamental feature concerning eukaryotic DING genes, rather than technical problems.

In the light of all these facts, it seems that an unknown number of proteins from human and from other eukaryotes are still to be discovered and characterized. These observations should stimulate studies to develop techniques to sequence heterochromatin and highly repeated regions in genomes and to understand function, features, and properties of unknown genes and proteins, that is, eukaryotic *DING* genes. In that way, the study of proteins with unknown nucleotidic and peptidic sequence could be easier in case of crystallization success. Thus, combined use of X-ray crystallography and mass spectrometry could be a good alternative to obtain complete and exact peptidic sequence of large proteins.

CONCLUSION

This study reports 38-kDa HPBP ab initio sequencing using an original approach combining the forces of two complementary techniques: X-ray crystallography and mass spectrometry. The very large amount of experimen-

tal work that was required in crystallography and mass spectrometry to establish the total and unambiguous amino acid sequence of this protein illustrate how useful were the genome sequence programs. It also shows that in some cases, mass spectrometry has the potential to sequence large proteins using several complementary MS/MS techniques, and confirms how useful information deduced from X-ray structure can be. The resulting HPBP sequence, corresponding to the first complete sequence of a eukaryotic DING protein family, will be helpful to study the recently discovered HPBP, and allow us to correct the three-dimensional model of HPBP. This sequence also constitutes an important data to understand the particular genetic behavior of DING protein family. These eukaryotic proteins, intriguingly systematically absent from nucleotidic database, might not be the only ones. As eukaryotic genomes are generally said to lack heterochromatin and highly repeated sequences, it seems that an unknown number of proteins from human or different organisms are still to be discovered. For these unknown proteins, if crystals diffracting at appropriate resolution are available, the combined use of techniques described in this article can provide the complete amino acids sequence of proteins, with no particular size limitation.

MATERIALS AND METHODS

Protein purification

The HPBP/HPON1-containing fraction was obtained by using the HPON1 purification protocol previously described.[34] Briefly, plasma bags (Etablissement Français du Sang Rhône-Alpes, Beynost, France) were submitted to a pseudo-affinity chromatography on Cibacron Blue (Sigma, L"Isle-d"Abeau, France) allowing the isolation of hydrophobic plasma proteins, mainly lipoproteins. Then the HPBP/HPON1-containing fraction was separated from all of the other HDL-bound proteins, mainly apoA-I, using Triton X-100 and DEAE anion exchange chromatography (Pharmacia Biotech, Uppsala, Sweden). The pure HPBP fraction was obtained using purification protocol from Renault *et al.*[18] HPBP/HPON1-containing fraction in 25 m*M* Tris buffer containing 0.1% Triton X-100, were injected on Bio-Gel HTP hydroxyapatite (BioRad Laboratories, Munich, Germany) equilibrated with 10 m*M* sodium phosphate pH 7.0. This step was followed by washing with the same buffer and elution by 400 m*M* sodium phosphate allowed to separate the two proteins. HPBP was not retained on hydroxyapatite equilibrated without $CaCl_2$ and was collected in the filtrate. On the contrary, HPON1 was retained and subsequently eluted by means of higher phosphate concentrations. These pure HPBP fractions were used for mass spectrometry experiments.

SDS-PAGE electrophoresis

Protein fractions were analyzed by SDS/PAGE performed according to the discontinuous system of Laemmli,[35] using the Bio-Rad Mini protean III electrophoresis unit. The polyacrylamide concentration was 10% (w/v) for the separating gel and 4% for the stacking gel. Prior to loading, samples were incubated in sample buffer containing 2% (w/v) SDS and 10% (w/v) glycerol, and heated for 5 min at 90°C. Prestained molecular weight markers (Dual color, Bio-rad) were loaded on each gel. Runs were carried out at a constant voltage (200 V) for 45 min. Gels were stained using Coomassie blue. Bands containing the pure 38-kDa HPBP were cut and used for in-gel digestion.

Crystallographically predicted sequence data

Amino acid sequence has been first attributed from electronic density map. Indeed, the shape of visible electronic density allow, in most of cases, to attribute the nature of the residue.

The complete amino acid sequence was first determined from X-ray data. This sequence was confirmed at 75% by classical methods, such as N-terminal sequencing, mass spectrometry experiments, and internal peptide digestion. These complementary analyses show an error rate of only 11% in the crystallographically predicted sequence. So, HPBP sequence provided by Morales *et al.*[9] is about 97% exact. We started from the three-dimensional model of HPBP and from the corresponding structure factors deposited in the Protein Data Bank (PDB id: 2cap). The complete and unambiguous sequence attribution of HPBP is reported in this article.

Three-dimensional model checking, manual building, and refinement

Mass spectrometry experiments described in the article allow us to correct HPBP sequence. These sequence correction have permitted to increase the three-dimensional model's (2cap) quality. Visualization of the model and manual corrections were performed using Coot,[36] and refinements were performed using *REFMAC*.[37] Stereo and ball and sticks representation of HPBP structure were made using *PyMOL*.[38]

ESI-MS on intact HPBP

Since HPBP was solubilized for purification in 0.1% Triton X-100, the detergent was removed before ESI-MS experiment. First an Extracti-Gel D Detergent Removing Gel kit (Pierce Biotechnology, Rockford, IL) was used, but the ESI-MS spectra obtained showed that large amounts of detergent were still present, which prevented an accurate mass measurement. Then, the remaining detergent was successfully removed using a C18 HPLC

chromatography. The protein was eluted at 60% acetonitrile in water (0.1% formic acid).

ESI-MS measurements were performed on an electrospray quadrupole time-of-flight mass spectrometer (Q-TOF II, Waters, Manchester, UK). Mass spectra were recorded in the positive ion mode on the mass range 500–2500 *m/z*, after calibration with horse heart myoglobin diluted to 2 pmol/μL in a water/acetonitrile mixture (1:1, v/v) acidified with 1% formic acid (v/v).

In-gel digestion

In-gel digestion was performed with an automated protein digestion system, MassPrep Station (Waters, Manchester, United Kingdom). The gel plugs were washed twice with 50 μL of 25 m*M* ammonium hydrogen carbonate (NH_4HCO_3) and 50 μL of acetonitrile. The cysteine residues were reduced by 50 μL of 10 m*M* dithiothreitol at 57°C and alkylated by 50 μL of 55 m*M* iodoacetamide. After dehydration with acetonitrile, the proteins were cleaved in gel with a solution of 12.5 ng/μL of modified porcine trypsin (Promega, Madison, WI) or endoproteinase Lys-C sequencing grade (Roche Diagnostics Corporation, Indianapolis, IN) or Chymotrypsin sequencing grade (Roche Diagnostics Corporation, Indianapolis, IN) in 25 m*M* NH_4HCO_3 or a solution of 10 ng/μL of thermolysin from *Bacillus thermoproteolyticus rokko* (Sigma-Aldrich, St. Louis, MO). The digestion was performed overnight at room temperature or 2 h at 65°C for thermolysin digestion. The generated peptides were extracted with 60% acetonitrile in 5% formic acid.

MALDI-MS and MALDI-MS/MS

MS and MS/MS spectra were acquired with the Ultraflex™ TOF/TOF mass spectrometer (Bruker Daltonics GmbH, Bremen, Germany) with gridless ion optics under control of Flexcontrol 2.0. This instrument equipped with the SCOUT™ High Resolution Optics with X-Y multisample probe and gridless reflector was used at a maximum accelerating potential of 25 kV and was operated in reflector mode. Ionization was accomplished with a 337-nm beam from a nitrogen laser with a repetition rate of 20 Hz. The output signal from the detector was digitized at a sampling rate of 2 GHz. The samples were prepared by standard dried droplet preparation on stainless steel MALDI targets using α-cyano-4-hydroxycinnamic acid as matrix.

The external calibration of MALDI mass spectra was carried out using singly charged monoisotopic peaks of a mixture of bradykinin 1–7 (m/z = 757.400), human angiotensin II (m/z = 1046.542), human angiotensin I (m/z = 1296.685), substance P (m/z = 1347.735), bombesin (m/z = 1619.822), renin (m/z = 1758.933), ACTH 1–17 (m/z = 2093.087), and ACTH 18–39 (m/z = 2465.199). To achieve mass accuracy, internal calibration was performed with tryptic peptides coming from autolysis of trypsin, with respectively monoisotopic masses at m/z = 842.510, m/z = 1045.564, and m/z = 2211.105. Monoisotopic peptide masses were automatically annotated using Flexanalysis 2.0 software.

The MALDI-MS/MS spectra were obtained by the analysis of the metastable ions, generated by Laser-Induced Decomposition (LID) of the selected precusor ions. No additional collision gas was applied. Precusor ions were accelerated to 8 kV and selected in a timed ion gate. The fragments were further accelerated by 19 kV in the LIFT cell and their masses were analyzed after the ion reflector passage. MALDI-MS/MS spectra were annotated with the Biotools 2.2 software package.

NanoLC-MS/MS on peptide digests

NanoLC-MS/MS analysis was performed either using a CapLC capillary LC system (Waters, Manchester, United Kingdom), coupled to a hybrid Quadrupole Time-Of-Flight mass spectrometer (Q-TOF 2, Waters, Manchester, United Kingdom).

From each sample, 6.4 μL was loaded on a precolumn, before chromatographic separation on a C18 column (LC Packings C18, 75 mm id, 150-mm length). The gradient was generated by the CapLC at a flow rate of 200 nL/min. The gradient profile consisted of a linear one from 90% of a water solution acidified by 0.1% formic acid (v/v) (solution A), to 40% of a solution of CH_3CN acidified by 0.1% formic acid (v/v) (solution B) in 30 min, followed by a second gradient ramp to 75% of B in 1 min. Data acquisition was piloted by MassLynx software V4.0. Calibration was performed using adducts of 0.1% phosphoric acid (Acros, NJ) with a scan range from m/z 50 to 1800. Automatic switching between MS and MS/MS modes was used. The internal parameters of Q-TOF II were set as follow. The electrospray capillary voltage was set to 3.5 kV, the cone voltage set to 35 V, and the source temperature set to 90°C. The MS survey scan was m/z 300–1500 with a scan time of 1 s and an interscan time of 0.1 s. When the peak intensity rose above a threshold of 15 counts/s, tandem mass spectra were acquired. Normalized collision energies for peptide fragmentation were set using the charge-state recognition files for 1+, 2+, and 3+ peptide ions. The scan range for MS/MS acquisition was from m/z 50 to 2000 with a scan time of 1 s and an interscan time of 0.1 s. Fragmentation was performed using argon as collision gas and with a collision energy profile optimized for various mass ranges of precursor ions. Data processing was done automatically with the ProteinLynx Process module.

Strategy for sequence correction

First, from the nanoLC-MS/MS data, a search was done against the last version of HPBP sequence using

Mascot (MatrixScience, Boston) software. Then, from the nonidentified peptides, which have a good quality of MS/MS spectra, de novo sequencing was performed, using Peaks (Bioinformatics Solutions Inc., Waterloo, Canada) Masslynx PepSeq (Waters, Manchester, United Kingdom) and Biotools Rapid de novo (Bruker Daltonics GmbH, Bremen, Germany) softwares. A tag is identified and a sequence proposed. The peptide corresponding to the tag was searched in HPBP sequence. Finally, from the sequence proposed by the software and the mass difference between the measured and theoretical masses, the errors could be identified and corrected.

ACKNOWLEDGMENTS

The Region Alsace is acknowledged for financing the MALDI Ultraflex instrument. D.R. is under contract with the German Bundesministerium der Verteidigung (M/SAB 1/6/A002).

REFERENCES

1. Berna A, Bernier F, Scott K, Stuhlmuller B. Ring up the curtain on DING proteins. FEBS Lett 2002;524:6–10.
2. Riah O, Dousset JC, Bofill-Cardona E, Courriere P. Isolation and microsequencing of a novel cotinine receptor. Cell Mol Neurobiol 2000;20:653–664.
3. Belenky M, Prasain J, Kim H, Barnes S. DING, a genistein target in human breast cancer: a protein without a gene. J Nutr 2003;133(7 Suppl):2497S–2501S.
4. Blass S, Schumann F, Hain NA, Engel JM, Stuhlmuller B, Burmester GR. p205 is a major target of autoreactive T cells in rheumatoid arthritis. Arthritis Rheum 1999;42:971–980.
5. Kumar V, Yu S, Farell G, Toback FG, Lieske JC. Renal epithelial cells constitutively produce a protein that blocks adhesion of crystals to their surface. Am J Physiol 2004;287:F373–F383.
6. Adams L, Davey S, Scott K. The DING protein: an autocrine growth-stimulatory protein related to the human synovial stimulatory protein. Biochim Biophys Acta 2002;1586:254–264.
7. Weebadda WK, Hoover GJ, Hunter DB, Hayes MA. Avian air sac and plasma proteins that bind surface polysaccharides of Escherichia coli O2. Comp Biochem Physiol 2001;130:299–312.
8. Scott K, Wu L. Functional properties of a recombinant bacterial DING protein: comparison with a homologous human protein. Biochimica Biophys Acta 2005;1744:234–244.
9. Morales R, Berna A, Carpentier P, Contreras-Martel C, Renault F, Nicodeme M, Chesne-Seck ML, Bernier F, Dupuy J, Schaeffer C, Diemer H, Van-Dorsselaer A, Fontecilla-Camps JC, Masson P, Rochu D, Chabriere E. Serendipitous discovery and X-ray structure of a human phosphate binding apolipoprotein. Structure 2006;14: 601–609.
10. Du M, Zhao L, Li C, Zhao G, Hu X. Purification and characterization of a novel fungi Se-containing protein from Se-enriched *Ganoderma lucidum* mushroom and its Se-dependent radical scavenging activity. Eur Food Res Technol 2007;224:659–665.
11. Chen Z, Franco CF, Baptista RP, Cabral JM, Coelho AV, Rodrigues CJ, Jr, Melo EP. Purification and identification of cutinases from *Colletotrichum kahawae* and Colletotrichum gloeosporioides. Appl Microbiol Biotechnol 2007;73:1306–1313.
12. Berna A, Bernier F, Chabriere E, Perera T, Scott K. DING proteins: novel members of a prokaryotic phosphate-binding protein superfamily which extends into the eukaryotic kingdom. Intl J Biochem Cell Biol 2007, http://dx.doi.org/10.1016/j.biocel.2007.02.004; doi: 10.1016/j.biocel.2007.02.004.
13. Mehta A, Lu X, Block T, Willis A, Dwek R, Tennant B, Blumberg B. Synovial stimulatory protein fragments copurify with woodchuck hepatitis virus: implications for the etiology of arthritis in chronic hepatitis B virus infection. Arthritis Rheum 2001;44:486–487.
14. Hain N, Alsalameh S, Bertling WM, Kalden JR, Burmester GR. Stimulation of rheumatoid synovial and blood T cells and lines by synovial fluid and interleukin-2: characterization of clones and recognition of a co-stimulatory effect. Rheumatol Int 1990;10:203–210.
15. Contreras-Martel C, Carpentier P, Morales R, Renault F, Chesne-Seck ML, Rochu D, Masson P, Fontecilla-Camps JC, Chabriere E. Crystallization and preliminary X-ray diffraction analysis of human phosphate-binding protein. Acta Crystallogr F Struct Biol Cryst Commun 2006;62 (Part 1):67–69.
16. Fokine A, Morales R, Contreras-Martel C, Carpentier P, Renault F, Rochu D, Chabriere E. Direct phasing at low resolution of a protein copurified with human paraoxonase (PON1). Acta Crystallogr 2003; 59 (Part 12):2083–2087.
17. Hain NA, Stuhlmuller B, Hahn GR, Kalden JR, Deutzmann R, Burmester GR. Biochemical characterization and microsequencing of a 205-kDa synovial protein stimulatory for T cells and reactive with rheumatoid factor containing sera. J Immunol 1996;157:1773–1780.
18. Renault F, Chabriere E, Andrieu JP, Dublet B, Masson P, Rochu D. Tandem purification of two HDL-associated partner proteins in human plasma, paraoxonase (PON1) and phosphate binding protein (HPBP) using hydroxyapatite chromatography. J Chromatogr 2006; 836:15–21.
19. Webb MR. A tale of the unexpected. Structure 2006;14:391–392.
20. Chao BH, Jakubowski JA, Savage B, Chow EP, Marzec UM, Harker LA, Maraganore JM. Agkistrodon piscivorus piscivorus platelet aggregation inhibitor: a potent inhibitor of platelet activation. Proc Natl Acad Sci USA 1989;86:8050–8054.
21. Kostanjevecki V, Leys D, Van Driessche G, Meyer TE, Cusanovich MA, Fischer U, Guisez Y, Van Beeumen J. Structure and characterization of Ectothiorhodospira vacuolata cytochrome b(558), a prokaryotic homologue of cytochrome b(5). J Biol Chem 1999;274: 35614–35620.
22. Hellman U, Eggertsen G, Engstrom A, Sjoquist J. Amino acid sequence of the trypsin-generated C3d fragment from human complement factor C3. Biochem J 1985;230:353–361.
23. Parthasarathy S, Murthy MR. Analysis of temperature factor distribution in high-resolution protein structures. Protein Sci 1997;6: 2561–2567.
24. Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins 2005;58:905–912.
25. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V. Improving gene annotation using peptide mass spectrometry. Genome Res 2007;17:231–239.
26. Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics (Oxford, England) 2005;21:4322–4329.
27. Galindo MI, Pueyo JI, Fouix S, Bishop SA, Couso JP. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. PLoS Biol 2007;5:e106.
28. Perrodou E, Deshayes C, Muller J, Schaeffer C, Van Dorsselaer A, Ripp R, Poch O, Reyrat JM, Lecompte O. ICDS database: interrupted CoDing sequences in prokaryotic genomes. Nucleic Acids Res 2006;34:D338–D343.
29. Deshayes C, Perrodou E, Gallien S, Euphrasie D, Schaeffer C, Van-Dorsselaer A, Poch O, Lecompte O, Reyrat JM. Interrupted coding sequences in *Mycobacterium smegmatis*: authentic mutations or sequencing errors? Genome Biol 2007;8:R20.
30. Gade D, Theiss D, Lange D, Mirgorodskaya E, Lombardot T, Glockner FO, Kube M, Reinhardt R, Amann R, Lehrach H, Rabus R, Gobom J. Towards the proteome of the marine bacterium *Rhodopir-*

ellula baltica: mapping the soluble proteins. Proteomics 2005;5: 3654–3671.
31. Riano-Pachon DM, Dreyer I, Mueller-Roeber B. Orphan transcripts in *Arabidopsis thaliana*: identification of several hundred previously unrecognized genes. Plant J 2005;43:205–212.
32. International Human Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 2004;432: 695–716.
33. de Diego JG, David Rodriguez F, Rodriguez Lorenzo JL, Grappin P, Cervantes E. cDNA-AFLP analysis of seed germination in *Arabidopsis thaliana* identifies transposons and new genomic sequences. J Plant Physiol 2006;163:452–462.
34. Gan KN, Smolen A, Eckerson HW, La Du BN. Purification of human serum paraoxonase/arylesterase. Evidence for one esterase catalyzing both activities. Drug Metab Dispos: Biol Fate Chem 1991;19:100–106.
35. Laemmli U. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 1970;227:6.
36. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr 2004;60 (Part 12, 1):2126–2132.
37. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr 1997;53 (Part 3):240–255.
38. DeLano WL. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific; 2002.