Journal of Computer-Aided Molecular Design, **15:** 309–322, 2001.
KLUWER/ESCOM
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

The search for a new model structure of β-Factor XIIa

Elsa S. Henriques[a], Welly B. Floriano[b], Nathalie Reuter[c], André Melo[a], David Brown[d], José A.N.F. Gomes[a], Bernard Maigret[c], Marco A.C. Nascimento[e] & Maria João Ramos[a,*]
[a]*CEQUP/Departamento de Química, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal;* [b]*Centro de Ciencias Exatas, Departamento de Fisica, Universidade Federal do Espirito Santo, Av. Fernando Ferrari s/n, Goiabeiras, Vitoria ES 29060-900, Brazil;* [c]*Laboratoire de Chimie Theorique UMR CNRS 7565, Université de Nancy I, Domaine Scientifique Victor Grignard B.P. 239, 54506 Vandoeuvre les Nancy Cedex, France;* [d]*LMOPS, Bât. IUT, Université de Savoie, Campus Scientifique, Savoie Technolac, 73376 Le Bourget du Lac, France;* [e]*Instituto de Quimica, Departamento de Fisico-Quimica, Cidade Universitaria, CT Bloco A sala 412, Rio de Janeiro RJ 21949-900, Brazil*

Received 29 February 2000; accepted 18 December 2000

Key words: buried water molecules, β-factor XIIa, homology modelling, molecular dynamics, molecular modelling, protein structure, serine protease

Summary

We present the search for a new model of β-factor XIIa, a blood coagulation enzyme, with an unknown experimental 3D-structure. We decided to build not one but three different models using different homologous proteins as well as different techniques and different modellers. Additional studies, including extensive molecular dynamics simulations on the solvated state, allowed us to draw several conclusions concerning homology modelling, in general, and β-factor XIIa, in particular.

Introduction

When a blood vessel injury occurs in a healthy mammalian circulatory system, blood coagulation immediately takes place in order to protect its integrity. The process of blood coagulation is extremely complex consisting of a series of interactions involving, amongst others, 13 or 14 plasma glycoproteins all of which are zymogens of serine proteases [1, 2].

One of the blood coagulation agents is the Hageman factor, a name also attributed to human factor XII, a glycoprotein with a molecular weight of 80 000 Da that has been previously purified and characterised [3, 4]. Factor XII is transformed into α-factor XIIa when in the presence of kallikrein and a negatively charged surface, yielding β-factor XIIa upon further proteolysis. The amino acid sequences of the two proteolytic products of activated factor XII have been established [5, 6], and confirmed later by the determination of the organisation of the human factor XII gene [7]. Both α-factor XIIa and β-factor XIIa have been reported to consist of two polypeptide chains each, a light chain (28 000 and 2000 Da of molecular weight for α- and β-factor XIIa, respectively) and a heavy chain (52 000 and 28 000 Da of molecular weight for α- and β-factor XIIa respectively) held together by a disulphide bond. As far as β-factor XIIa is concerned, the light chain, the L-chain, contains 9 amino acids and the heavy chain, the H-chain, is composed of 243 amino acids. Besides the disulphide bond that holds both the L- and H-chains together, β-factor XIIa has six additional internal disulphide bonds.

Despite all the information presently available on β-factor XIIa, its structure is still experimentally unknown as, indeed, is also that of α-factor XIIa. Computer models have been attempted before [8, 9], and are the only means of visualisation for experimentalists in need of this structure. At that time, only a few experimental crystal structures belonging to the large serine protease family and homologous to β-factor

[*] To whom correspondence should be addressed. E-mail: mjramos@fc.up.pt

Table 1. Databank search, sequence alignment, modelling and refinement details for the computer model A.

DATABANK SEARCH		
OWL Databank [11] using programs SWEEP [12] and BLAST [13]		
HOMOLOGOUS PROTEINS USED FOR THE SEQUENCE ALIGNMENT		
PDB code	Description	% ID[a]
2kai	pancreatic kallikrein A (pig)	34
1mct	pancreatic trypsin (pig)	36
1ntp	pancreatic β-trypsin (modified) (bovine)	36
1bra	trypsin (variant D189G, G226D) (black rat)	34
2tbs	trypsin (salmon)	34
1est	pancreatic tosyl-elastase (pig)	27
SEQUENCE ALIGNMENT		
Selected proteins and β-factor XIIa were aligned using program CLUSTAL W [14]. Secondary structure assignment for pdb files was made using PROCHECK [15] and included in the alignment. Additionally, a secondary structure prediction [16] was made for β-factor XIIa and also taken into account		
MODELLING		
Model obtained using SWISS–MODEL [17]		
REFINEMENT		
The model solvated with 920 TIP 3P waster molecules was minimised initially using the CHARMM force field [18], and subsequently the AMBER [20] force field		

[a] % ID stands for percentage of sequence identity. The latter was calculated according to the number of sequence matches in the alignment of a protein with β-factor XIIa (100% ID); this results in different ID percentages for the same protein according to the particular alignment they are subjected to.

XIIa were available. Nowadays, the number of structurally known proteins has increased tremendously, and modelling techniques have greatly advanced in recent years. Being extremely interested in the design of new inhibitors for β-factor XIIa, and having found a few discrepancies in the existing models, we have decided to tackle this problem by producing a new, more reliable computer model of the enzyme in the first place. In fact, we have decided to achieve this by building, not one, but three models using different homologous enzymes, different homology techniques, and also different modellers. This technique has been used at least in one very famous exercise [10], and we have decided to adopt it to try and increase our chances of success as, in the present case, we are dealing with sequence identities that are relatively low (∼40%). We have learned far more about β-factor XIIa by doing this than we would have done by building only one model. In the end, we were able to compare the models, and subsequently obtain the best model to use in future studies. Extensive molecular dynamics simulation studies have yielded interesting results, which seem to point to one particular direction.

Materials and methods

Data bank search, sequence alignment and modelling

This section is further divided into three subsequent parts, each of which relates the details of the sequence alignments as well as the modelling schemes, and structure refinements of the three models (A, B and C) built for β-factor XIIa. The numbering used throughout the article is sequencial within each chain, L- and H-, unless when it is specifically referred to a particular pdb structure; in those cases, the numbering of the different amino acids follows the one used in the correspondent sequence and/or pdb files.

Model A

The serine protease domain of Hageman factor ranging from Val_{373} to Ser_{615} corresponds to the H-chain of β-factor XIIa. This sequence of amino acids was used to search the OWL databank [11] using the programs SWEEP [12] and BLAST [13]. The results were analysed and the proteins selected, with 3D structure known and available, showed some workable degree of homology to the referred protease domain of human factor XII. This enzyme has a preference for an arginine residue in its specificity pocket, and we used this

```
KLK_PIG_PD      --------------------------IIGGRECEKNSHPWQVAIYH--YSSFQ--CGGVL
NRL_1MCTA       --------------------------IVGGYTCAANSIPYQVSLNS--GSHF---CGGSL
NRL_1NTP__      --------------------------IVGGYTCGANTVPYQVSLNS--GYHF---CGGSL
NRL_1BRA__      --------------------------IVGGYTCQENSVPYQVSLNS--GYHF---CGGSL
NRL_2TBS        --------------------------IVGGYECKAYSQAHQVSLNS--GYHF---CGGSL
FA12_HUMAN      --------------------------VVGGLVALRGAHPYIAALYW--GHSF---CAGSL
EL1_PIG__P      MLRLLVVASLVLYGHSTQDFPETNARVVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTL
                                          ..**  .   .     ..           * * *

KLK_PIG_PD      VNPKWVLTAAHCKNDN-----YEVWLGRHNLFENENTAQFFGVTADFPHPGFNLSADGKD
NRL_1MCTA       INSQWVVSAAHCYKSR-----IQVRLGEHNIDVLEGNEQFINAAKIITHPNF----NGNT
NRL_1NTP__      IDSQWVVSAAHCYKSG-----IQVRLGEDNINVVEGNEQFISASKSIVHPSY----DSNT
NRL_1BRA__      INDQWVVSAAHCYKSR-----IQVRLGEHNINVLEGNEQFVNAAKIIKHPNF----DRKT
NRL_2TBS        VNENWVVSAAHCYKSR-----VEVRLGEHNIKVTEGSEQFISSSRVIRHPNY----SSYN
FA12_HUMAN      IAPCWVLTAAHCLQDRPAPEDLTVVLGQERRNHSCEPCQTLAVRSYRLHEAF----SPVS
EL1_PIG__P      IRQNWVMTAAHCVDRELT---FRVVVGEHNLNQNDGTEQYVGVQKIVVHPYWN--TDDVA
                .   **..****           * .*           *         *

KLK_PIG_PD      YSHDLMLLRLQSPAK-----ITDAVKVLELPTQEP--ELGSTCEASGWGSIEPGPDDFEF
NRL_1MCTA       LDNDIMLIKLSSPAT-----LNSRVATVSLPRSCA--AAGTECLISGWGNTK--SSGSSY
NRL_1NTP__      LNNDIMLIKLKSAAS-----LDSRVASISLPTSCA--SAGTQCLISGWGNTK--SSGTSY
NRL_1BRA__      LNNDIMLIKLSSPVK-----LNARVATVALPSSCA--PAGTQCLISGWGNTL--SSGVNE
NRL_2TBS        IDNDIMLIKLSKPAT-----LNTYVQPVALPTSCA--PAGTMCTVSGWGNTM--SSTAD-
FA12_HUMAN      YQHDLALLRLQEDADGSCALLSPYVQPVCLPSGAARPSETTLCQVAGWGHQF--EGAEEY
EL1_PIG__P      AGYDIALLRLAQSVT-----LNSYVQLGVLPRAGTILANNSPCYITGWGLTR---TNGQL
                   *. *..*          .   *    **         . *  .***

KLK_PIG_PD      PDEIQCVQLTLLQNTFC--ADAHPDKVTESMLCAGYLPGGKDTCMGDSGGPLICNG----
NRL_1MCTA       PSLLQCLKAPVLSNSSC--KSSYPGQITGNMICVGFLQGGKDSCQGDSGGPVVCNG----
NRL_1NTP__      PDVLKCLKAPILSDSSC--KSAYPGQITSNMFCAGYLEGGKDSCQGDSGGPVVCSG----
NRL_1BRA__      PDLLQCLDAPLLPQADC--EASYPGKITDNMVCVGFLEGGKGSCQGDSGGPVVCNG----
NRL_2TBS        SDKLQCLNIPILSYSDC--NDSYPGMITNAMFCAGYLEGGKDSCQGDSGGPVVCNG----
FA12_HUMAN      ASFLQEAQVPFLSLERCSAPDVHGSSILPGMLCAGFLEGGTDACQGDSGGPLVCEDQAAE
EL1_PIG__P      AQTLQQAYLPTVDYAICSSSSYWGSTVKNSMVCAGGD-GVRSGCQGDSGGPLHCLVNG--
                   ..      .    *         .   * * *   *    * ******. *

KLK_PIG_PD      ---MWQGITSWGHTP-CGSANKPSIYTKLIFYLDWIDDTITENP
NRL_1MCTA       ---QLQGIVSWGYG--CAQKNKPGVYTKVCNYVNWIQQTIAAN-
NRL_1NTP__      ---KLQGIVSWGSG--CAQKNKPGVYTKVCNYVSWIKQTIASN-
NRL_1BRA__      ---ELQGIVSWGYG--CALPDNPDVYTKVCNYVDWIQDTIAAN-
NRL_2TBS        ---ELQGVVSWGYG--CAEPGNPGVYAKVCIFSDWLTSTMASY-
FA12_HUMAN      RRLTLQGIISWGSG--CGDRNKPGVYTDVAYYLAWIREHTVS--
EL1_PIG__P      -QYAVHGVTSFVSRLGCNVTRKPTVFTRVSAYISWINNVIASN-
                     .*. *      *     * ... .  .  *.
```

RED	β-sheet
CYAN	helix
BLUE	catalytic triad
MAGENTA	specificity pocket
GREEN	oxyanion hole
RED	predicted β-sheet
CYAN	predicted helix

Figure 1. CLUSTAL W (1.60) [14] multiple sequence alignment for model A of β-factor XIIa. Regions in red are identified with β-sheets and helices are in cyan; the predicted β-sheets and helices obtained using PREDICTPROTEIN [16] have been underlined additionally. Also shown are the catalytic triad in blue, the specificity pocket in magenta and the oxyanion hole in green.

fact as an additional selection criterion – the availability of inhibited 3D structures having an arginine type inhibitor. The final list of homologous proteins, along with other relevant information, is shown in Table 1.

The selected proteins and β-factor XIIa were aligned using the CLUSTAL W program [14]. The secondary structure assignment for the pdb files was made using PROCHECK [15] and included in the alignment. Additionally, a secondary structure prediction [16] was made for β-factor XIIa and also taken into account. The resulting alignment is shown in Figure 1 along with the catalytic triad, the oxyanion hole and the aspartic acid present into the specificity pocket, all of which are characteristic of the serine protease trypsin-like family of proteins. An assignment of secondary structures was made based on the final alignment, and a list of matched amino acids from the homologous proteins was drawn for β-factor XIIa.

The actual modelling process of β-factor XIIa was initiated by generating a first model using SWISS-MODEL [17], an automated protein modelling server. The model was based on the pdb entries *1mct*, *1npt*, *1bra*, *2kai* and *1est*. Arginine, lysine, aspartate and glutamate residues were considered as charged. Chlorine and sodium ions were added to neutralise charges. A primary refinement of the model was made using the programme CHARMm [18]. This primary model was subsequently refined, and its secondary structure assignment was improved to match the one based on our alignment. The seven disulphide bridges known to be present in β-factor XIIa were verified. Four of them are conserved in the other enzymes and therefore easy to assign (residues 26–42, 129–198, 161–177 and 188–219). One of the bridges connects the two chains (H- and L-) of β-factor XIIa together and therefore could not be formed. The two remaining disulphide bridges can be assigned based on the cys-cys distance (residues 65–680), SWISS-PROT [19] information and previous models [8, 9] (residues 34–104). The model solvated with 920 TIP3P water molecules was then minimized using AMBER [20].

Model B

We have used the SWISS-PROT sequence (FA12_HU MAN, P00748) of the human factor XII precursor (EC 3.4.21.38), and extracted from this sequence the amino acids ranging from Val_{373} to Ser_{615} corresponding to the H-chain of β-factor XIIa. This sequence of amino acids was subsequently used to search for similarities with other protein sequences, using the EMBL databank and the FASTA [21] program. The results showed that the best homology was reached with three protein types: hepatocyte growth factor activator, tissue plasminogen activator (t-PA) and urokinase-type plasminogen activator (u-PA). X-ray structures were made available for the latter two proteins (*1rtf* and *1lmw* pdb codes respectively). The sequence alignment between β-factor XIIa and the protease part of these two plasminogen activators is presented in Figure 2.

This alignment was further used in the homology modelling. The HOMOLOGY program of the BIOSYM [22] modelling package was used for that purpose. Using the alignment proposed above, this program assigned coordinates for the β-factor XIIa parts, which had been set to be homologous with parts of the *1rtf* and *1lmw* 3D structures. The sidechains were positioned according to their most suitable similarities, with the corresponding amino acids in the target proteins. The loop searching procedure proposed in the HOMOLOGY program was used for adding the missing parts. The crude model of β-factor XIIa thus obtained was next refined using several rounds of energy minimisation procedures and the CVFF force field in the DISCOVER BIOSYM [22] program. In the first step the backbone of the regions of β-factor XIIa having a good similarity with those of *1rtf* and *1lmw* proteins were frozen. Next, all constraints were removed and the structure fully energy refined until convergence; the conjugate gradient algorithm was used for that purpose. No cut-off was set up in these calculations, and the dielectric constant was chosen as distance dependent. The N- and C-terminal groups as well as the arginine, lysine, aspartate and glutamate residues were considered as being charged. Table 2 presents the information relative to the building of model B.

Model C

Upon a thorough examination of the more recently released crystallographic structures [23, 24] of various members of the serine protease S1 family, we selected a set of possible templates for a new model of human β-factor XIIa. Both sequence and secondary structure homologies were weighed in the amino acid sequence alignments, and the final list of homologous proteins as well as the respective percentages of sequence identity are shown in Table 3.

Unlike models A and B, model C includes both the L- and the H-chains of β-factor XIIa. However, the percentages of sequence identity mentioned are exclusively related to the H-chain for easy comparison with

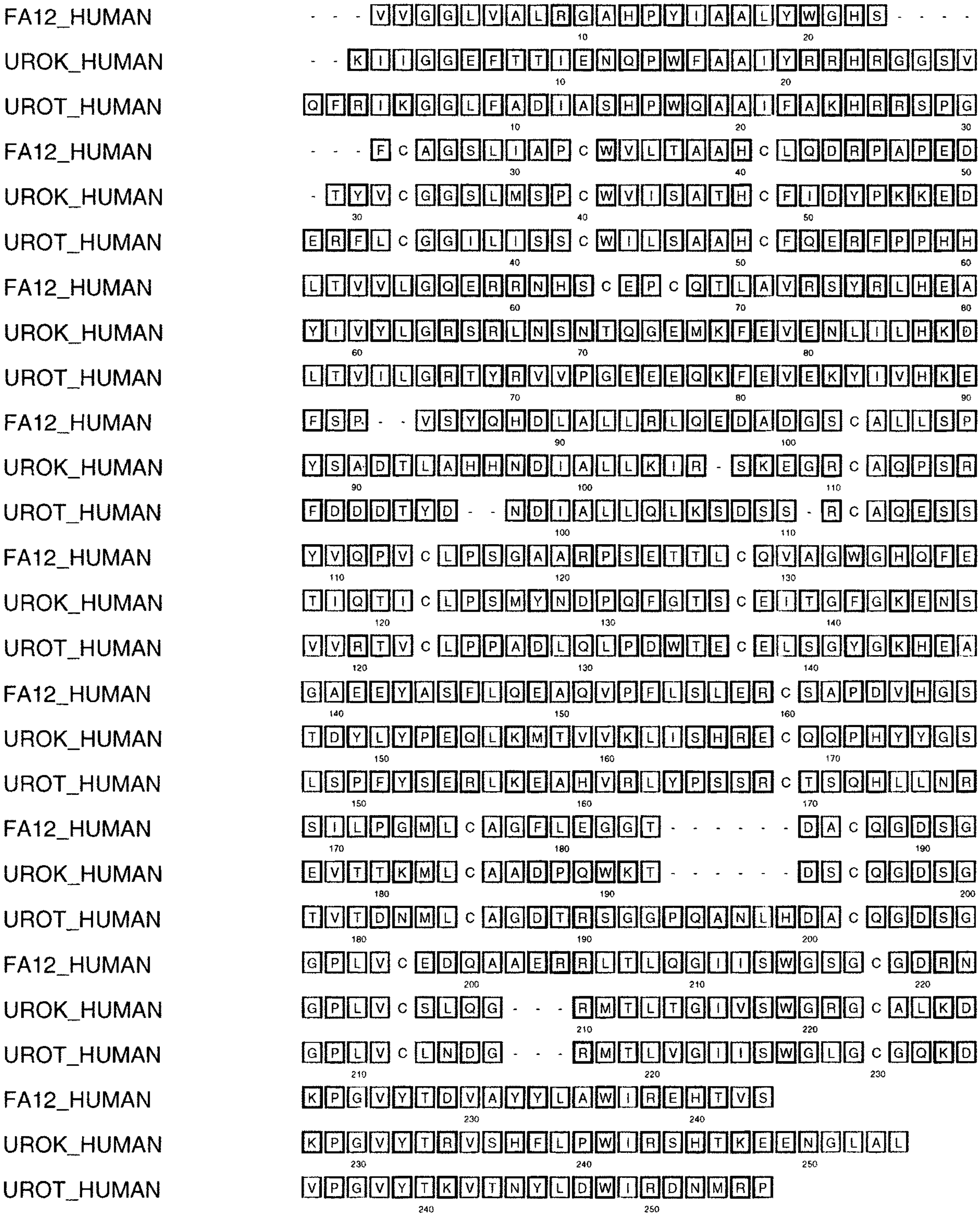

FA12_HUMAN - - - V V G G L V A L R G A H P Y I A A L Y W G H S - - - -
10 20
UROK_HUMAN - - K I I G G E F T T I E N Q P W F A A I Y R R H R G G S V
10 20
UROT_HUMAN Q F R I K G G L F A D I A S H P W Q A A I F A K H R R S P G
10 20 30
FA12_HUMAN - - - F C A G S L I A P C W V L T A A H C L Q D R P A P E D
30 40 50
UROK_HUMAN - T Y V C G G S L M S P C W V I S A T H C F I D Y P K K E D
30 40 50
UROT_HUMAN E R F L C G G I L I S S C W I L S A A H C F Q E R F P P H H
40 50 60
FA12_HUMAN L T V V L G Q E R R N H S C E P C Q T L A V R S Y R L H E A
60 70 80
UROK_HUMAN Y I V Y L G R S R L N S N T Q G E M K F E V E N L I L H K D
60 70 80
UROT_HUMAN L T V I L G R T Y R V V P G E E E Q K F E V E K Y I V H K E
70 80 90
FA12_HUMAN F S P - - V S Y Q H D L A L L R L Q E D A D G S C A L L S P
90 100
UROK_HUMAN Y S A D T L A H H N D I A L L K I R - S K E G R C A Q P S R
90 100 110
UROT_HUMAN F D D D T Y D - - N D I A L L Q L K S D S S - R C A Q E S S
100 110
FA12_HUMAN Y V Q P V C L P S G A A R P S E T T L C Q V A G W G H Q F E
110 120 130
UROK_HUMAN T I Q T I C L P S M Y N D P Q F G T S C E I T G F G K E N S
120 130 140
UROT_HUMAN V V R T V C L P P A D L Q L P D W T E C E L S G Y G K H E A
120 130 140
FA12_HUMAN G A E E Y A S F L Q E A Q V P F L S L E R C S A P D V H G S
140 150 160
UROK_HUMAN T D Y L Y P E Q L K M T V V K L I S H R E C Q Q P H Y Y G S
150 160 170
UROT_HUMAN L S P F Y S E R L K E A H V R L Y P S S R C T S Q H L L N R
150 160 170
FA12_HUMAN S I L P G M L C A G F L E G G T - - - - - - D A C Q G D S G
170 180 190
UROK_HUMAN E V T T K M L C A A D P Q W K T - - - - - - D S C Q G D S G
180 190 200
UROT_HUMAN T V T D N M L C A G D T R S G G P Q A N L H D A C Q G D S G
180 190 200
FA12_HUMAN G P L V C E D Q A A E R R L T L Q G I I S W G S G C G D R N
200 210 220
UROK_HUMAN G P L V C S L Q G - - - R M T L T G I V S W G R G C A L K D
210 220
UROT_HUMAN G P L V C L N D G - - - R M T L V G I I S W G L G C G Q K D
210 220 230
FA12_HUMAN K P G V Y T D V A Y Y L A W I R E H T V S
230 240
UROK_HUMAN K P G V Y T R V S H F L P W I R S H T K E E N G L A L
230 240 250
UROT_HUMAN V P G V Y T K V T N Y L D W I R D N M R P
240 250

Figure 2. β-factor XIIa H-chain sequence alignment to the two template enzymes urokinase-type and tissue plasminogen activators for model B.

Table 2. Databank search, sequence alignment, modelling and refinement details for computer model B.

DATABANK SEARCH		
EMBL databank using program FASTA [21]		
HOMOLOGOUS PROTEINS USED FOR THE SEQUENCE ALIGNMENT		
PDB code	Description	% ID[a]
1rtf	Tc-tissue plasminogen activator (human)	41
1lmw	lmW-urokinase-type plasminogen activator (human)	39
SEQUENCE ALIGNMENT		
Selected proteins and β-factor XIIa were aligned using HOMOLOGY from BIOSYM [22] modelling package		
MODELLING		
Model obtained using HOMOLOGY from BIOSYM [22] modelling package		
REFINEMENT		
Refinement done using DISCOVER from BIOSYM [22] modelling package		

[a] % ID stands for percentage of sequence identity. The latter was calculated according to the number of sequence matches in the alignment of a protein with β-factor XIIa (100% ID); this results in different ID percentages for the same protein according to the particular alignment they are subjected to.

Table 3. Databank search, sequence alignment, modelling and refinement details for computer model C.

DATABANK SEARCH		
Protein Data Bank [23]		
Homologous proteins used for the sequence alignment		
PDB code	Description	% ID[a]
1rtf	Tc-tissue plasminogen activator (human)	42
1lmw	lmW-urokinase-type plasminogen activator (human)	40
4ptp	pancreatic β-trypsin (bovine)	36
2kai	pancreatic kallikrein A (pig)	35
1nes	pancreatic ε-elastase (pig)	33
SEQUENCE ALIGNMENT		
Selected proteins and β-factor XIIa were aligned manually by sequencial and structural homology		
MODELLING		
Model obtained using QUANTA [26]		
Introduction of a set of conserved buried waters known to be preserved in enzymes sharing the specificity of trypsin with positions predetermined by structural homology [27, 30]		
REFINEMENT		
Refinement done using CHARMM [18]		

[a] % ID stands for percentage of sequence identity. The latter was calculated according to the number of sequence matches in the alignment of a protein with β-factor XIIa (100% ID); this results in different ID percentages for the same protein according to the particular alignment they are subjected to.

the ones obtained with models A and B. The alignment was done manually by sequencial and structural homology, using relevant data from the literature [8, 9, 25]. The final sequence alignment is depicted in Figure 3.

Placement of indels is one of the factors that affects models the most. For example, in Figure 2, the first Asp following the gap is part of the catalytic triad. This is typical of what can be obtained from a primary sequence alignment. On the other hand, the alignment shown in Figure 3 is much closer to a structural alignment. This is one of the factors that could explain some of the differences between models B and C.

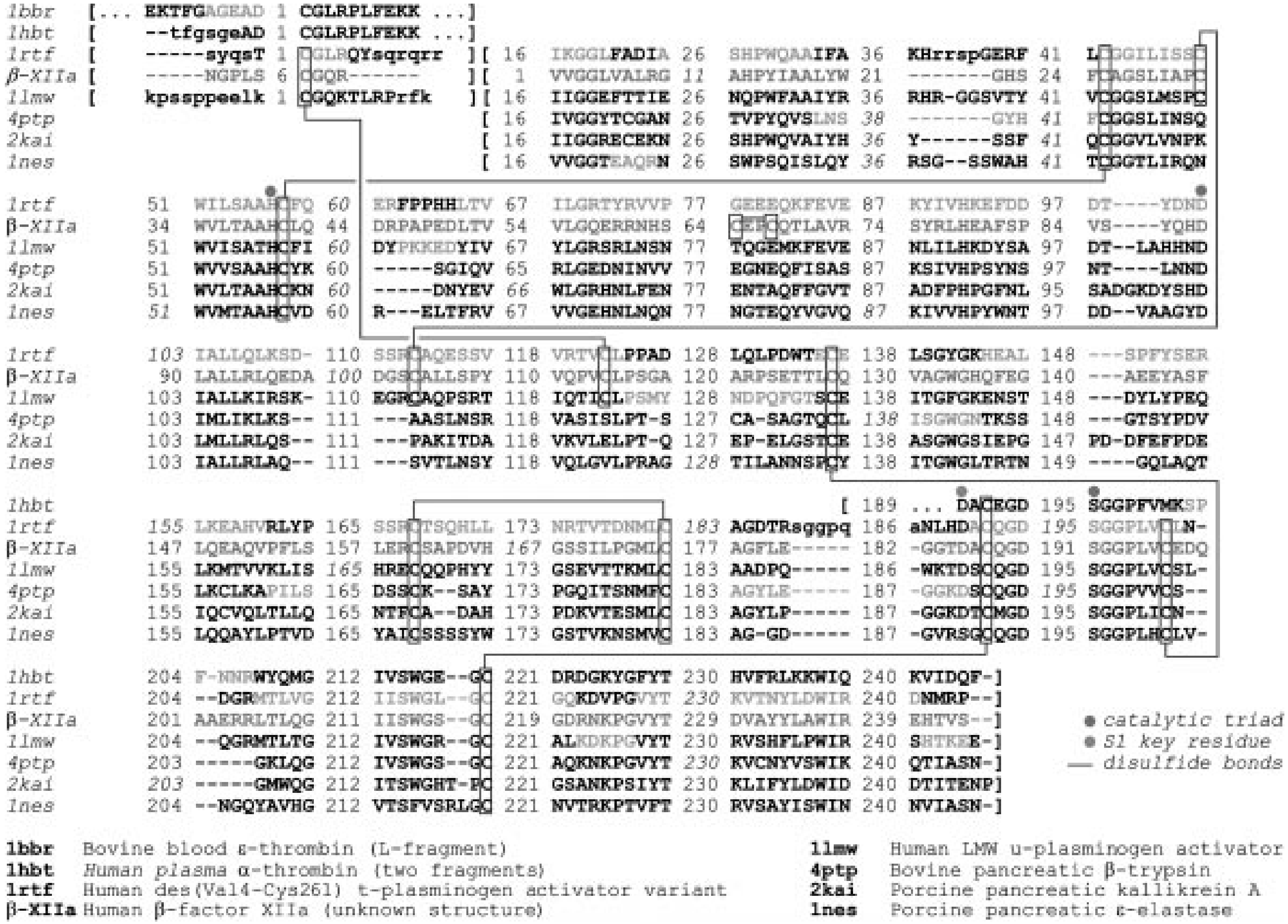

1bbr [... EKTFGAGEAD 1 CGLRPLFEKK ...]
1hbt [--tfgsgeAD 1 CGLRPLFEKK ...]
1rtf [-----syqsT 1 CGLRQYsqrqrr][16 IKGGLFADIA 26 SHPWQAAIFA 36 KHrrspGERF 41 LCGGILISSC
β-XIIa [-----NGPLS 6 CGQR------][1 VVGGLVALRG 11 AHPYIAALYW 21 -------GHS 24 FCAGSLIAPC
1lmw [kpssppeelk 1 CGQKTLRPrfk][16 IIGGEFTTIE 26 NQPWFAAIYR 36 RHR-GGSVTY 41 VCGGSLMSPC
4ptp [16 IVGGYTCGAN 26 TVPYQVSLNS 38 -------GYH 41 FCGGSLINSQ
2kai [16 IIGGRECEKN 26 SHPWQVAIYH 36 Y------SSF 41 QCGGVLVNPK
1nes [16 VVGGTEAQRN 26 SWPSQISLQY 36 RSG--SSWAH 41 TCGGTLIRQN
1rtf 51 WILSAAHCFQ 60 ERFPPHHLTV 67 ILGRTYRVVP 77 GEEEQKFEVE 87 KYIVHKEFDD 97 DT----YDND
β-XIIa 34 WVLTAAHCLQ 44 DRPAPEDLTV 54 VLGQERRNHS 64 CEPCQTLAVR 74 SYRLHEAFSP 84 VS----YQHD
1lmw 51 WVISATHCFI 60 DYPKKEDYIV 67 YLGRSRLNSN 77 TQGEMKFEVE 87 NLILHKDYSA 97 DT--LAHHND
4ptp 51 WVVSAAHCYK 60 -----SGIQV 65 RLGEDNINVV 77 EGNEQFISAS 87 KSIVHPSYNS 97 NT----LNND
2kai 51 WVLTAAHCKN 60 -----DNYEV 66 WLGRHNLFEN 77 ENTAQFFGVT 87 ADFPHPGFNL 95 SADGKDYSHD
1nes 51 WVMTAAHCVD 60 R---ELTFRV 67 VVGEHNLNQN 77 NGTEQYVGVQ 87 KIVVHPYWNT 97 DD--VAAGYD
1rtf 103 IALLQLKSD- 110 SSRCAQESSV 118 VRTVCLPPAD 128 LQLPDWTECE 138 LSGYGKHEAL 148 ---SPFYSER
β-XIIa 90 LALLRLQEDA 100 DGSCALLSPY 110 VQPVCLPSGA 120 ARPSETTLCQ 130 VAGWGHQFEG 140 ---AEEYASF
1lmw 103 IALLKIRSK- 110 EGRCAQPSRT 118 IQTICLPSMY 128 NDPQFGTSCE 138 ITGFGKENST 148 ---DYLYPEQ
4ptp 103 IMLIKLKS-- 111 ---AASLNSR 118 VASISLPT-S 127 CA-SAGTQCL 138 ISGWGNTKSS 148 ---GTSYPDV
2kai 103 LMLLRLQS-- 111 ---PAKITDA 118 VKVLELPT-Q 127 EP-ELGSTCE 138 ASGWGSIEPG 147 PD-DFEFPDE
1nes 103 IALLRLAQ-- 111 ---SVTLNSY 118 VQLGVLPRAG 128 TILANNSPCY 138 ITGWGLTRTN 149 ----GQLAQT
1hbt [189 ... DACEGD 195 SGGPFVMKSP
1rtf 155 LKEAHVRLYP 165 SSRCTSQHLL 173 NRTVTDNMLC 183 AGDTRsggpq 186 aNLHDACQGD 195 SGGPLVCLN-
β-XIIa 147 LQEAQVPFLS 157 LERCSAPDVH 167 GSSILPGMLC 177 AGFLE----- 182 -GGTDACQGD 191 SGGPLVCEDQ
1lmw 155 LKMTVVKLIS 165 HRECQQPHYY 173 GSEVTTKMLC 183 AADPQ----- 186 -WKTDSCQGD 195 SGGPLVCSL-
4ptp 155 LKCLKAPILS 165 DSSCK--SAY 173 PGQITSNMFC 183 AGYLE----- 187 -GGKDSCQGD 195 SGGPVVCS--
2kai 155 IQCVQLTLLQ 165 NTFCA--DAH 173 PDKVTESMLC 183 AGYLP----- 187 -GGKDTCMGD 195 SGGPLICN--
1nes 155 LQQAYLPTVD 165 YAICSSSSYW 173 GSTVKNSMVC 183 AG-GD----- 187 -GVRSGCQGD 195 SGGPLHCLV-
1hbt 204 F-NNRWYQMG 212 IVSWGE--GC 221 DRDGKYGFYT 230 HVFRLKKWIQ 240 KVIDQF-]
1rtf 204 --DGRMTLVG 212 IISWGL--GC 221 GQKDVPGVYT 230 KVTNYLDWIR 240 DNMRP--]
β-XIIa 201 AAERRLTLQG 211 IISWGS--GC 219 GDRNKPGVYT 229 DVAYYLAWIR 239 EHTVS--]
1lmw 204 --QGRMTLTG 212 IVSWGR--GC 221 ALKDKPGVYT 230 RVSHFLPWIR 240 SHTKEE-]
4ptp 203 -----GKLQG 212 IVSWGS--GC 221 AQKNKPGVYT 230 KVCNYVSWIK 240 QTIASN-]
2kai 203 -----GMWQG 212 ITSWGHT-PC 221 GSANKPSIYT 230 KLIFYLDWID 240 DTITENP]
1nes 204 --NGQYAVHG 212 VTSFVSRLGC 221 NVTRKPTVFT 230 RVSAYISWIN 240 NVIASN-]
● catalytic triad
● S1 key residue
— disulfide bonds
1bbr Bovine blood ε-thrombin (L-fragment)
1hbt Human plasma α-thrombin (two fragments)
1rtf Human des(Val4-Cys261) t-plasminogen activator variant
β-XIIa Human β-factor XIIa (unknown structure)
1lmw Human LMW u-plasminogen activator
4ptp Bovine pancreatic β-trypsin
2kai Porcine pancreatic kallikrein A
1nes Porcine pancreatic ε-elastase

Figure 3. Multiple sequence alignment of β-factor XIIa for model C. Both L- and H-chains are considered. Gaps (–) were introduced to optimize sequence alignments. Modelled residues are in pink, and template residues in light blue. The catalytic triad is signaled red, the S1 key residue is marked green and the seven disulphide bridges have been outlined in blue. Residues in small letters are not defined in the respective pdb file.

The modelling of β-factor XIIa was carried out using the software QUANTA [26]; Figure 3 shows which parts, of which 3D protein structure, have been used to model its own structure. CHARMm generated the residues with no correspondent ones in the templates (residues Ala_{109} and Ala_{202}), as well as all the hydrogens.

The prediction of the new three-dimensional structure of β-factor XIIa was further improved by the introduction of a set of conserved buried waters, known to be preserved in enzymes sharing the primary specificity of trypsin [27], with positions predetermined by structural homology. We obtained a coordinate matrix of the conserved buried waters, resulting from a comparative study of their protein environments to those in the templates and trypsin-like structures. This matrix, shown in Table 4, was added to the modelled structure of β-factor XIIa and the water hydrogen positions were generated using CHARMm.

The resulting model of β-factor XIIa was finally energy minimised using CHARMm. The backbone was frozen and harmonic constraints were imposed on all the oxygens of the conserved buried waters; the system went through enough minimisation steps to correct all the bad contacts. A cut-off distance of 15 Å and a distance dependent dielectric constant were used. On the whole, neither any significant alterations of the water positions nor any abnormal distortions of the sidechains were observed. All the relevant information concerning the building of model C is shown in Table 3.

Table 4. Coordinate matrix of the oxygen atoms belonging to the 22 conserved buried waters (BW), resulting from a comparative study of their protein environments to those in human tissue plasminogen activator (*1rtf*), bovine pancreatic β-trypsin (*4ptp*), and bovine pancreatic β-trypsin with bovine pancreatic trypsin inhibitor, BPTI (*2ptc*). Residues in bold were added to the modelled structure of β-factor XIIa, and the water hydrogen positions were generated using CHARMM [18].

BW-site	1rtf	4ptp	2ptc
1(296)	**DSOL:15**	2H:326	
2(297)	**DSOL:9**	2H:314	
3(298)	–	**2H:358**	
4(299)	**DSOL:26**	2H:353	
5(300)	**DSOL:24**	2H:352	
6(301)	**DSOL:12**	2H:331	
7(302)	**DSOL:14**	2H:275	
8(303)	**DSOL:7**	2H:328	
9(304)	–	**2H:307**	
10(305)	**DSOL:13**	2H:330	
11(306)	–	**2H:340**	
12(307)	–	**2H:334**	
13(308)	–	**2H:339**	
14(309)	**DSOL:17**	2H:349	3H:416
15(310)	**DSOL:18**	2H:260	
16(311)	**DSOL:19**	2H:354	
17(312)	**DSOL:11**	2H:347	
18(313)	**DSOL:43**	2H:360	
19(314)	**DSOL:6**	2H:346	
20(315)	–	–	**3H:541**
21(316)	**DSOL:31**	2H:269	
22(317)	**DSOL:16**	2H:345	3H:414

Note: As pointed out in the text, the conformation of the backbone of residues 22 to 27 is different in *1rtf* and *1lmw,* relatively to other structures, being associated with waters 9, 11, 12 and 13.

Molecular dynamics

Finally, all three models were submitted to 1ns of molecular dynamics (MD) as follows: For each model, a set of MD simulations was carried out under periodic boundary conditions, using a cubic box of side length 80 Å. Water molecules were added in order to fill the box to a realistic density (1001.75 kg/m^3). The N- and C- terminal groups, and the Arg, Asp and Glu residues were charged and, therefore, Na^+ ions were added to neutralise the charges. The counter ions were placed at random at the border of the water box to avoid 'trapped' interactions with the protein. During the MD runs, none approached the protein to interact with a sidechain, which was the expected result. They kept their hydration shell and were moving in the bulk. The full system subjected to MD thus contained β-factor XIIa (3574 atoms), 12 Na^+ ions neutralising the charge on the system, and 15 682 water molecules, i.e., a total of 50 632 atoms altogether.

The entire system was subjected to MD simulations using the parallel program *ddgmq* [28] with the interaction potential based on the valence force field CVFF [22]. The potential used differs slightly from that used in CVFF in two respects. First, the angle bending is harmonic in the cosine of the angle, rather than in the angle itself, which requires a lesser computational effort in the calculations. Second, the out-of-plane term is harmonic in the distance of the central (trivalent) atom from the plane of the other three. This form was preferred over the improper dihedral form used in CVFF as the latter is somewhat ill-defined; there are three possible definitions of the improper dihedral angle and hence three different resulting energies. In both the angle bending and out-of-plane bending cases the corresponding force constant for use in *ddgmq* was obtained from the CVFF form by equating the curvatures at the minimum of the potentials.

After equilibration at 300 K for 20 ps, the system was allowed to relax under NPT conditions at an applied pressure of 1 bar. The simulation was then continued under NVT conditions using loose-coupling to a thermal bath and a coupling constant of 10 ps for another 800 ps, for a total simulation duration of 1000 ps. With the following Ewald summation parameters ($\alpha = 0.2$ Å^{-1}, $R_C = 11.5$ Å, $K_{max} = 10$), and a tolerance of 10^{-5} used in the SHAKE routine maintaining all bond lengths rigid, the program run in an Sgi Origin 2000. After the 1 ns MD simulation, the system was cooled to 1 K using MD before a final energy minimisation was performed.

Results and discussion

Homology modelling

The three-modelled structures of β-factor XIIa were homology built using different serine proteases as template structures, as well as different programs and different modellers. There was not any particular sort of strategy planned before the work was started and all the models were built simultaneously in time and with no pre-arranged restrictions, although in different laboratories.

Figure 4 depicts the 3D structures of all three models for β-factor XIIa. We have drawn it as a stereo figure, with the three models below each other in the same orientation for better clarity.

A detailed analysis of Tables 1–3 shows that model A was built based on kallikrein, elastase and various modified trypsins. On the other hand, with model B these enzymes have been completely discarded and only t-PA and u-PA have been considered. Finally, model C uses all of the enzymes that models A and B have considered with the exception of the various modified trypsins present in the former. All these choices have been explained in the previous section.

We think that the decision of using various modified trypsins for model A was a poor one, and has biased the final result because of the weight that the trypsin sequence has played in it. The reason of this choice was basically the attempt for using a prototype of the serine proteases family; trypsin is in a way the head of the family, an enzyme with a lower specificity which has been used again and again in all sorts of studies and models. On the other hand, the building of model B probably could have survived mainly on the two plasminogen activators if it was not for the fact that no experimental considerations were taken into account. These will be referred to next.

Most trypsin-like serine proteases feature an $\alpha/3_{10}$-helix centred around Cys_{168} (chymotrypsinogen numbering system [29]) and so do models A and B (top-left of each view in Figure 4); in t-PA, however, this helix is twisted terminating after one turn and carrying on two residues further ahead. If one follows the 'trypsin trend', residue His_{166} will be completely buried in the protein, just as it happens with model A and B; however, if t-PA is used for the modelling of this helix, as in model C, that same histidine will be exposed to the solvent, which seems to be the correct situation according to Ford et al.'s experimental work [30], concerning the binding of an antibody to β-factor XIIa. Therefore, in model C, the α-helix under discussion ends up being cut in two as shown in Figure 4.

Water molecules sequestered from bulk solvent within a protein matrix – buried waters – are integral conserved components of all serine proteases of known 3D structure [27]. In fact, Henriques et al. [31] have suggested that conserved buried waters should be included into any serine protease model built on the basis of sequence/structural homology to this family, since their absence might induce errors in a force field simulation, favouring the formation of non-existent hydrogen bonds, and locally inaccurate structure. The inclusion of the buried waters had a preponderant effect on the modelling of the loop constituted by residues 6–9, as it will be discussed next.

In model C, region 6–9 follows the *1nes/2kai* backbone, as opposed to u-PA/t-PA. These latter structures present a 'flipping' of the 18–19 peptide bond when compared with (chymo)trypsin, resulting in a deep burying of Ile_{24} in both proteins. β-factor XIIa features an arginine in the same correspondent position (residue 9 belonging to the H-chain); if that particular region is modelled using t-PA or u-TA, as in model B, the said arginine stays abnormally buried in the protein in an essentially hydrophobic cavity too small to accommodate its sidechain. However, the modelling using *1nes/2kai*, (which features a Lys/Arg in that particular position), as is done in model C, solves the problem and simultaneously allows for the inclusion

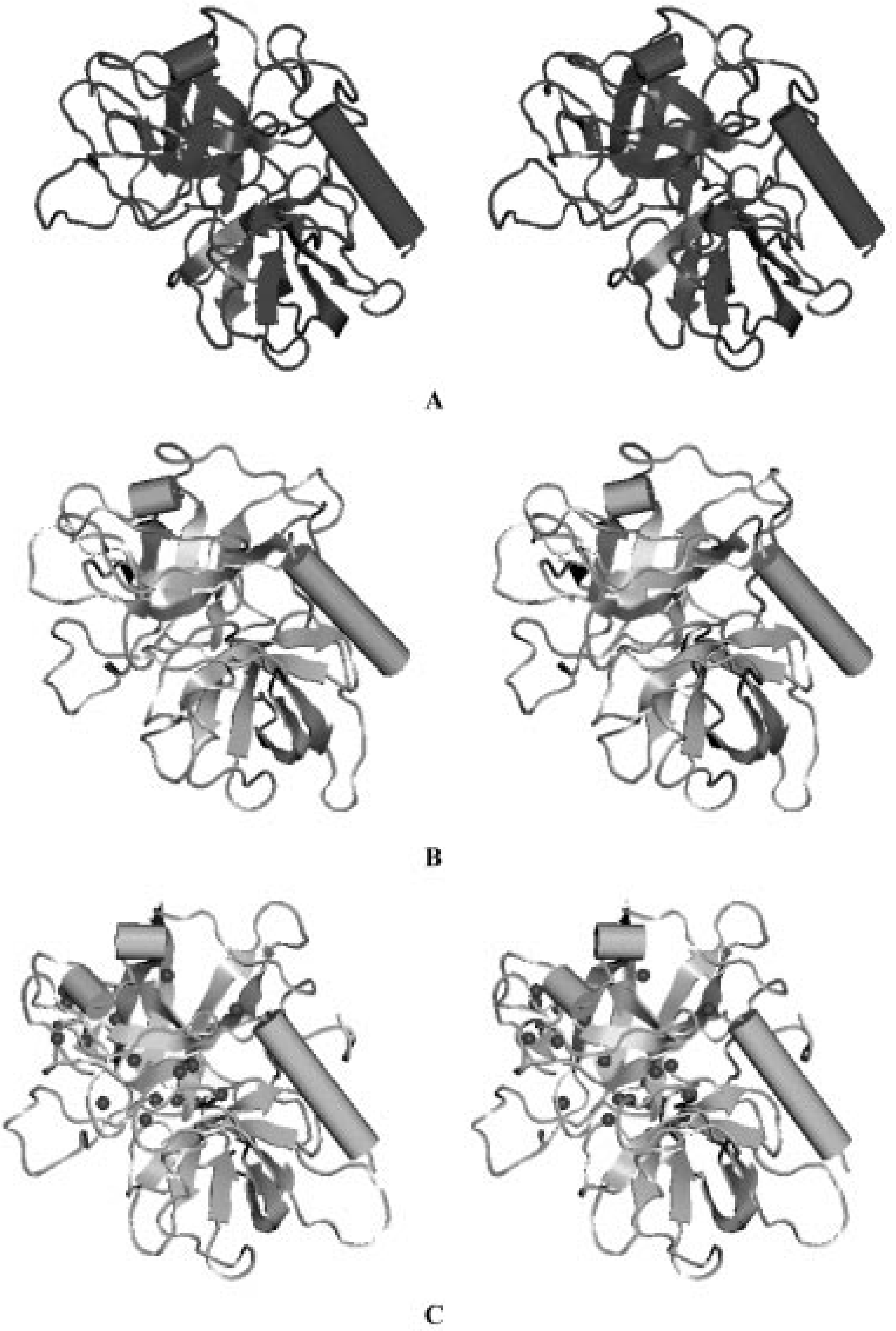


Figure 4. Stereo view of the 3D structures for models A, B and C - ribbon representation of the secondary structure. Only the H-chain of β-factor XIIa is presented, for comparison. Model C also shows the conserved buried waters.

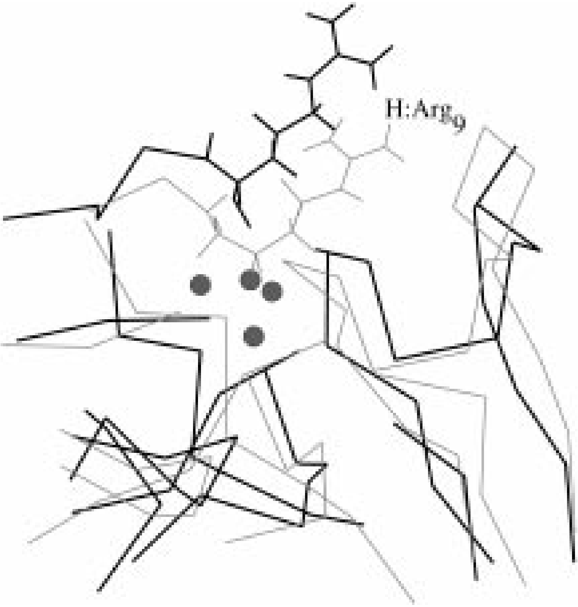


Figure 5. Backbone flip details for amino acid sequence region 6–9 in models B (grey) and C (black), with an Arg residue in position 9 of the H-chain and conserved buried waters 9, 11, 12 and 13 (as mentioned in Table IV).

of buried waters 9, 11, 12 and 13 (see Table 4). Figure 5 shows the backbone flip details for this particular region, as well as the Arg position and the concerned buried waters.

The inclusion of the L-chain is also worth mentioning as it corroborates what has just been explained in the previous paragraph; being much shorter in the carboxyl-end than both the corresponding ones in u-PA/t-PA, it should not protrude into region 6–9 (H-chain) as it seems to happen in both plasminogen activators. The modelling, using *1nes/2kai*, is in agreement with these thoughts.

Energy minimisation

The models were energy minimized prior to being submitted to molecular dynamics simulations. The details of each minimization procedure have been mentioned already in the text. After the energy minimization refinement, the quality of the resulting structures A, B, and C was also assessed with PROCHECK [15]. The main Ramachandran maps [32] and the corresponding plot statistics are shown in Figure 6. It can be readily observed that model C shows very good results whereas the other two models do not do so well.

Here we would like to mention the fact that full minimization could account for some of the differences between models B and C. As described in the previous section, the former model suffered full minimization as opposed to the latter one. In reality, as it has been pointed out before [33], energy minimization runs should be short to avoid the introduction of a large number of small errors.

Molecular dynamics

The way each structure evolved towards a stable one under the MD trajectories differs markedly. Figure 7 shows the RMS evolution of the two model structures B and C, according to their own starting conformation. Model A behaved in an erratic way, which basically confirmed the results obtained earlier with PROCHECK.

As can be observed from the figure, model C behaved rather well, having reached equilibrium during the simulation; model B did not do so well during that period of time. It seems reasonable to assume that the model which converges fastest under MD is likely to be of better quality. In fact, presently, a time-feasible refinement cannot move a large model significantly; this means that if a structure is far removed from the real one, convergence will not be possible. However, if convergence is quickly reached and even though we cannot infer from the fact that we are near reality, we should be able to at least assume that the model we have built has a better chance of being a good one.

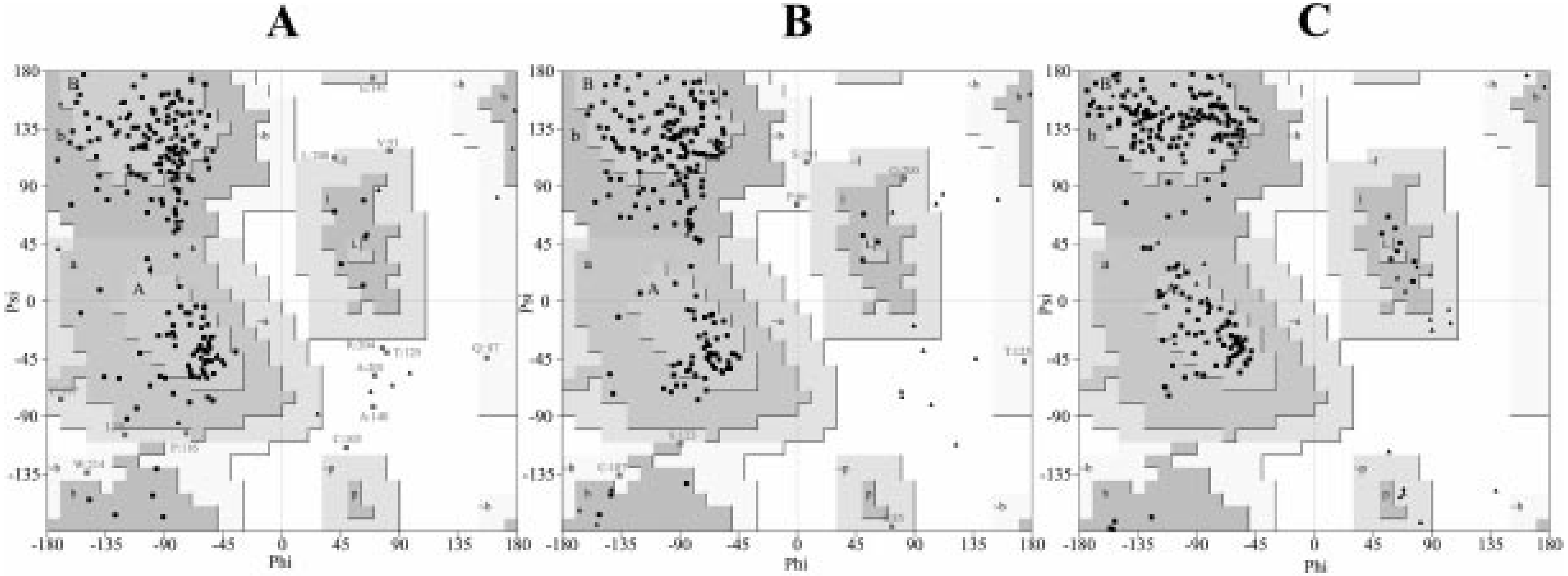
Figure 8 shows a picture of the adopted model (C) of β-factor XIIa, at the very end of the modelling.

Conclusion

The main conclusion of this article is obviously the fact that we believe that we now have a reliable model for β-factor XIIa, i.e., model C. Interestingly enough, it is the model built with most human intervention which proves to be the best. This model has better stereochemical parameters and, under refinement, has converged more quickly to a stable low RMS from its unrefined state.

Additionally, however, we did make several observations that might be of interest to people working in comparative molecular modelling. Some of them are well known to many even if not written down; others might be beneficial to some.

The exercise of building more than just one model for β-factor XIIa, using different approaches, proved to be extremely useful inasmuch as it forced us to examine quite different aspects of the problem.



Plot statistics

	A		B		C	
Residues in most favoured regions [A,B,L]	129	63.9%	142	70.3%	172	83.5%
Residues in additional allowed regions [a,b,l,p]	61	30.2%	54	26.7%	34	16.5%
Residues in generously allowed regions [~a,~b,~l,~p]	6	3.0%	6	3.0%	0	0.0%
Residues in disallowed regions	6	3.0%	0	0.0%	0	0.0%
	----	-------	----	-------	----	-------
Number of non-glycine and non-proline residues	202	100.0%	202	100.0%	206	100.0%
Number of end-residues (excl. Gly and Pro)	2		2		4	
Number of glycine residues (shown as triangles)	24		24		26	
Number of proline residues	15		15		16	
	----		----		----	
Total number of residues	243		243		252	

Figure 6. Ramachandran maps [31] and corresponding plot statistics for the energy minimized structures of β-factor XIIa (models A, B and C). Major ‘unacceptable’ Phi/psi values are signaled with the corresponding residue numbers.

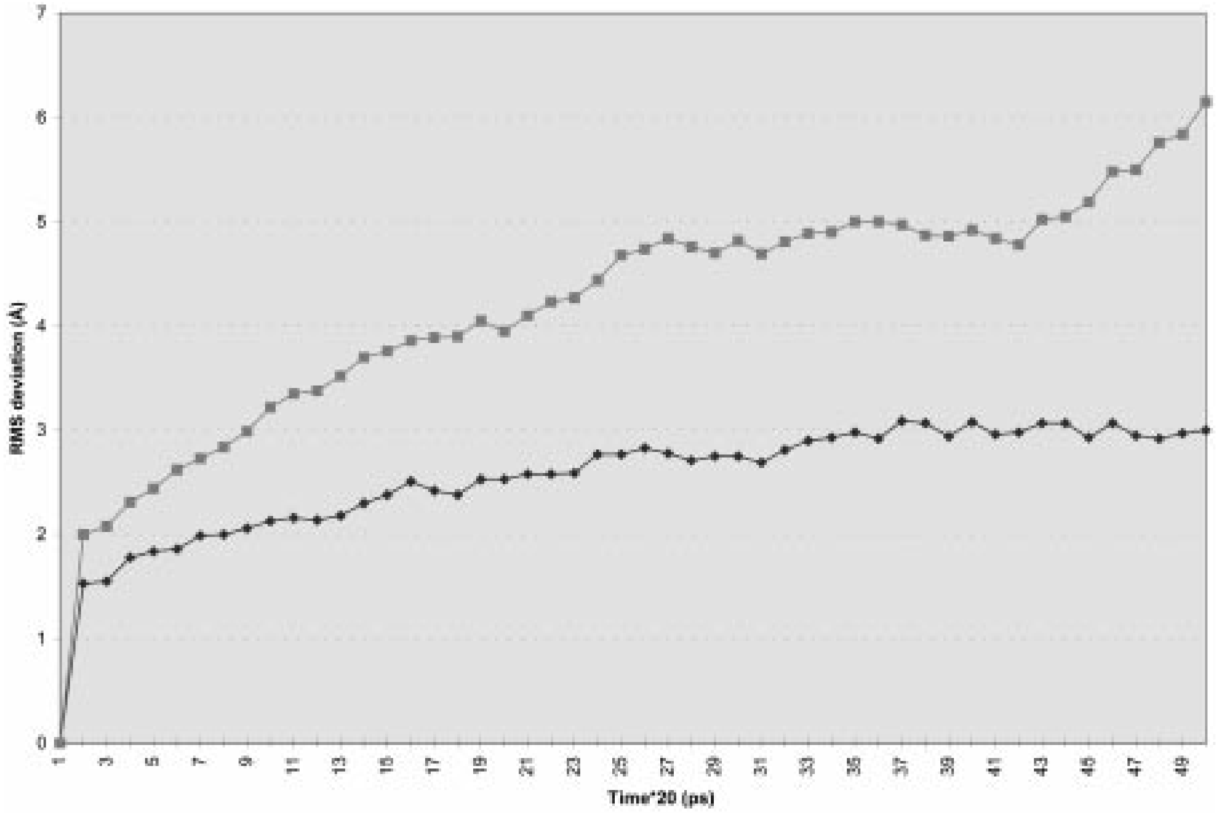


Figure 7. RMS (Å) variations of each model structure according to its own starting conformation. Line in magenta refers to model B and line in blue refers to model C.

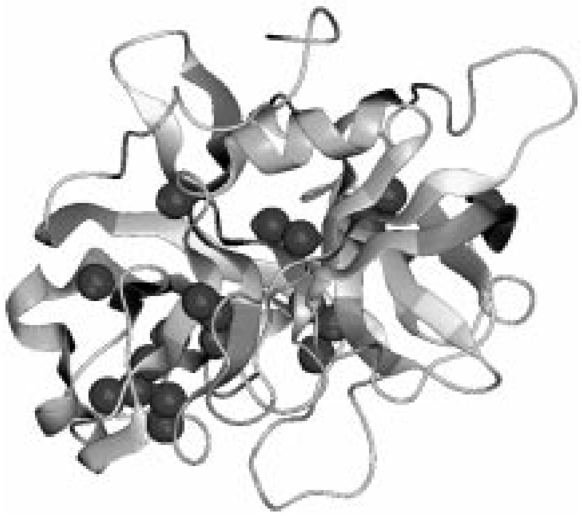


Figure 8. Final structure of the adopted model of β-factor XIIa, including the L-chain.

The inclusion of the whole protein (both chains in the present case) seems to be important; the reverse is a trap in which modellers often fall into, by looking just at the part of the protein containing the active centre and disregarding any other associated chains.

The main message of the work is that much attention should be put into analysing the modelling templates and the alignment used to build a model. The inclusion, for the sequence alignment, of several structures using mutations of the same protein seems to be inefficient whereas the use of as many different homologous structures as possible yielded good results. As far as the sequence alignment is concerned, it is important to check the results provided by the software, relying on human knowledge, intuition and common sense related to the particular case under study.

As far as the 3D modelling is concerned, when connecting two segments from two different templates one can easily introduce backbone conformational uncertainties which reflect later in bad Ramachandran values. It is, thus, rewarding to put quite a lot of effort into this part of the work.

Furthermore, the inclusion of conserved buried water molecules in the 3D model structures of serine proteases is crucial not only for force field simulation reasons, as stated previously, but also to help modelling locally accurate structure as it happened in this particular case.

One also ought to be careful with energy minimizations, performing short runs to avoid the introduction of a great number of small errors. In fact, we believe that this problem together with the inclusion of conserved buried waters and the careful connection of segments from different templates are probably the main causes which have led to much poorer Ramachandran values of model B in relation to model C.

Additionally, a good knowledge of the experimental behaviour of the protein to be modelled is obviously a very important asset.

All three models of β–factor XIIa are available upon request.

Acknowledgements

We thank the EU for research funding under project no. CI1*-CT94-006. We thank FCT (PRAXIS XXI) for a research grant for ESH.

References

1. Furie, B.and Furie, B.C., Cell, 53 (1988) 505.
2. Broze, G.J., Jr., Annu. Rev. Med., 46 (1995) 103.
3. Revak, S.D., Cochrane, C.G., Johnson, A.R. and Hugli, T.E., J. Clin. Invest., 54 (1974) 619.
4. Fujikawa, K. and Davie, E.W., Methods Enzymol., 80 (1981) 198.
5. McMullen, B.A. and Fujikawa, K., J. Biol. Chem., 260 (1985) 5328.
6. Fujikawa, K. and McMullen, B.A., J. Biol. Chem., 258 (1983) 10924.
7. Cool, D.E. and MacGillivray, R.T.A., J. Biol. Chem., 262 (1987) 13662.
8. Cool, D.E., G.V.Louie, C.-J.S., Zoller, M.J., Brayer, G.D. and MacGillivray, R.T.A., J. Biol. Chem., 260 (1985) 13 666.
9. Ramos, M.J., J. Mol. Graphics, 9 (1991) 91.
10. Mosimann, S., Meleshko, R. and James, M.N.G., Proteins: Structure, Function Genetics, 23 (1995) 301.
11. OWL: Bleasby, A.J. and Wootton, J.C., Prot. Engng, 3 (1990) 153.
12. Akrigg, D., Bleasby, A.J., Dix, N.I.M., Findlay, J.B.C., North, A.C.T., Parry-Smith, D.J., Wootton, J.C., Blundell, T.L., Gardner, S.P., Hayes, F., Islam, S., Sternberg, M.J.E., Thornton, J.M., Tickle, I.J. and Murray-Rust, P., Nature, 335 (1988) 745.
13. BLAST: Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., J. Mol. Biol., 215 (1990) 403.
14. CLUSTAL W: Thompson, J.D., Higgins, D.G. and Gibson, T.J., Nucleic Acids Res., 22 (1994) 4673.
15. PROCHECK v.2.1.4.: Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M., J. Appl. Cryst., 26 (1993) 283.
16. PREDICTPROTEIN: Rost, B., Meth. Enzym., 266 (1996) 525.
17. SWISS-MODEL (Automated Protein Modelling Server) ProMod: Peitsch, M.C., Biotechnology, 13 (1995) 658; Biochem. Soc. Trans., 24 (1996) 274.
18. CHARMm (release version c23f3), Molecular Simulations Incorporated.
19. SWISS-PROT: Appel, R.D., Bairoch, A. and Hochstrasser, D.F., Trends Biochem. Sci., 19 (1994) 258.
20. AMBER (version 4.1); Weiner, S.J. and Kollman, P.A., J. Am. Chem. Soc., 106 (1984) 765.
21. Pearson, W.R. and Lipman, D.J., Proc. Natl. Acad. Sci. USA, 85 (1988) 2444.
22. Biosym Technologies, 9685 Scranton Road, San Diego, CA (92121-4778).
23. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J., Protein Data Bank in Crystallographic Databases – Information Content, Software Systems, Scientific Applications, eds. F.H.Allen, G.Bergerhoff, R.Sievess, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, (1987) 107–132.
24. Protein Data Bank, Release # 75 (1996) Brookhaven National Laboratory, Upton, NY.
25. Lamba, D., Bauer, M., Huber, R., Fischer, S., Rudolph, R., Kohnert, U. and Bode, W., J. Mol. Biol., 258 (1996) 117.
26. QUANTA (version 4.0), 1994, Molecular Simulations Inc., 16 New England Executive Park, Burlington MA.
27. Sreenivasan, U. and Axelsen, P.H., Biochemistry, 31 (1992) 12 785.
28. Brown, D., Minoux, H. and Maigret, B., Comp. Phys. Commun., 103 (1997) 170.
29. Hartley, B.S. and Kauffman, D.L., Biochem. J., 101 (1966) 229.
30. Ford, R.P., Esnouf, M.P., Burgess, A.I. and Sarphie, A.F., J. Immunoassay, 17 (1996) 119.
31. Henriques, E.F., Ramos, M.J. and Reynolds, C.A., J. Comput. Aided Mol. Des., 11 (1997) 547.
32. Ramakrishnan, C. and Ramachandran, G.N., Biophys., 5 (1965) 909.
33. Rodriguez, R. and Vriend, G., website article at http://swift.embl-heidelberg.de/future/ articles/ text/ gambling.html (1998).