Journal of Biomedical Informatics xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Simulated pathogenic conformational switch regions matched well with the biochemical findings

Xin Liu, Ya-Pu Zhao *

The State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 1 July 2009
Available online xxxx

Keywords:
Conformational disease
Protein misfolding

ABSTRACT

Pathogenic conformational conversion is a general causation of many disease, such as transmissible spongiform encephalopathy (TSE) caused by misfolding of prion, sickle cell anemia, and etc. In such structural changes, misfolding occurs in regions important for the stability of native structure firstly. This destabilizes the normal conformation and leads to subsequent errors in folding pathway. Sites involved in the first stage can be deemed switch regions of the protein, and are vital for conformational conversion. Namely it could be a switch of disease at residue level. Here we report an algorithm that can identify such sites computationally with an accuracy of 93%, by calculating the probability of the native structure of a short segment jumping to a mistake one. Knowledge of such switch sites could be used to target clinical therapy, study physiological and pathologic mechanism of protein, and etc.

1. Introduction

Structure is the foundation of protein's role in metabolism. A misfolding of protein can induce an alteration in its biological function and properties, and result in conformation disease. A well-known case is prion disease, which is responsible of transmissible spongiform encephalopathy (TSE), a group of fatal neurodegenerative diseases in several mammalian species [1]. As a basis of life, protein takes part in nearly every biological process in life. Consequently, protein misfolding is a general causation of illness, and can result in various pathologic details. The scope of pathogenic conversional conversion is not limited to some classical well-known conformation diseases, but responsible for a large number of cases related to health.

Hydrophobic residues are tend to be blocked in the inner core of the native state of globular proteins [2], so that free energy of a system is minimized. As thus, protein structure is normally stable. An initial structural conversion usually occurs in the region significant for protein stability. It destabilizes the normal conformation and gives the opportunities to subsequent errors in folding pathway. Such initialization sites can be deemed switch regions of the protein, which face directly the key problem of misfolding—the origin of pathogenic structural conversion. If the initial misfolding can be controlled or prohibited, many physical process in the consequent misfolding will not happen. Therefore, investigations conducted on such regions would not give any chance for the unnecessary enlargement of the complicacy of research, such as misfolding pathway and the problems arising in subsequent refolding process. As vital for pathology, it should be significant in researches of corresponding diseases.

Another motivation of the present work is to optimize the knowledge contributed by clinical reports. Although many efforts have been made to uncover the nature of conformational disease, the large amount of sites reported by literature may confuse an inexpert, especially when some reports conflict with each other. The present independent judgment could lead a research to a pathway, which is suggested to be correct and significant in an aspect of physics. In any case, such insight is advantageous for the investigations in conformational diseases. Due to the drastic decrease of hardness and knowledge threshold, such work would be valuable particularly for interdisciplinary sciences.

Here we report an algorithm that can predict switch regions of pathogenic conformational changes using protein structural information, and affirm some significant sites in classical conformational disease regarding physics. It achieved successes in predicting diverse proteins that were believed to be responsible for conformational diseases. Both sensitivity and specificity are about 93%. The crucial sites for pathogenic structural change were successfully identified to be in a window about 15 residues. Only one tenth of the residues in test set were predicted to be vital for conformational conversion. Such high accuracy turns the algorithm into a practical tool of protein analysis.

2. Method

As the CDs come forth under the selection pressure of evolution, studying the start of disease-related misfolding is hard and a big

* Corresponding author. Address: The State Key Laboratory of Nonlinear Mechanics, Institute of Mechanics, Chinese Academy of Sciences, No. 15 Beisihuanxi Road, Beijing 100190, China. Fax: +86 010 82543977.
E-mail addresses: liuxin@lnm.imech.ac.cn (X. Liu), yzhao@imech.ac.cn (Y.-P. Zhao).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.12.003

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

challenge because the two tough problems—protein stability and protein evolution must be jointly considered. After hundreds of million years of evolution, many unstable protein mutants have been excluded by the selection pressure of evolution. Hardness of a proper representation of the selection pressure, together with stability feature, hampers the attempt. As thus, to the best of our knowledge, the special algorithm for this subject is still in absence.

A former work focusing on the topological feature of homologous relationships among polypeptides provides us an opportunity in solving the problem. In order to characterize the homologous relationships of short residue segments in the whole universe of non-membrane protein, we have presented a graph of polypeptide relationship (GPR) [3]. By sliding a 15-residue window along sequences, protein was treated as successive residue segments. Each 15-residue polypeptide served as a node of the graph. An edge was drawn between two nodes if the corresponding polypeptides are remote homologues. As thus, the GPR is a knowledge system derived from evolutionary information. We applied topological transitions to vital subgraphs of GRP by grouping homologous polypeptides together, and found that the phase space of polypeptide is composed of two nearly separated regions, a helix-donut zone and a strand-arc zone. Members of helix-donut zone are mainly helix segments and N- and C-terminal helix caps. The strand-arc is composed of β-sheet segments and N- and C-terminal strand caps. These two regions are sparsely connected by edges emanating from bridge nodes, shown on the right of Fig. 1A.

There are two traits in GPR. Firstly, the structures of the two zone are quite different. Once a polypeptide alters its state from one zone to another, there will be a structural change. Secondly, the selection pressure of evolution has been characterized by homologous relationship of the nodes of GPR. Therefore, the two obstacles for switch site prediction are overridden together by GPR. It provides a necessary foundation of present work.

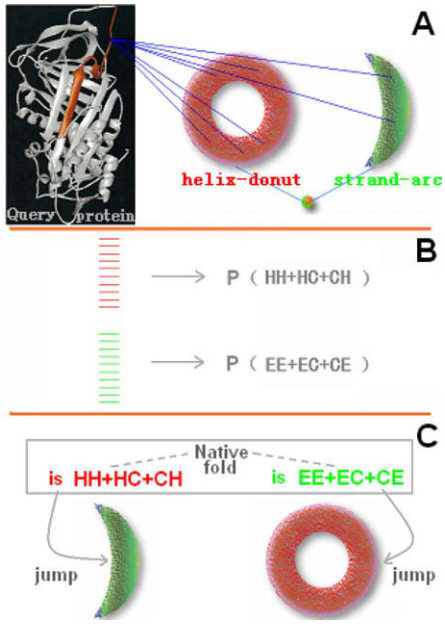

Fig. 1. Flow chart of the algorithm. As shown in the top-right graph, the whole universe of non-membrane residue segment contains two major regions: helix-donut zone (most members are helix segments and N/C-terminal helix caps) and strand-arc zone (mainly comprising β-sheet segments and N/C-terminal strand caps). (A) Identify remote-homologs of a query segment in the graph of polypeptide relationship (GPR). A query protein is treated as successive 15-residue segments. For each query segment, the remote-homologs in GPR are collected. Due to the obvious character in secondary structure, it is easy to indicate the region that a segment belongs to (with structural information). As thus, in step (B), the collected homologous segments are grouped according to their secondary structure (i.e. location in GPR; on the left of (B)), and then used to calculate the probabilities of the query segment belonging to the helix-donut zone and to the strand-arc zone. Such probabilities obtained in step B are then used to calculate the interchange probability of the query segment. As shown in (C), if the native state of a query segment is helix-donut/strand-arc, we calculate the probability in jumping to the strand-arc/helix-donut state.

In present work, we treat a query protein as successive 15-residue segments. For each query segment i, we identity its remote homologues in GPR using the method described in Section 3.1 of reference [3]. Nodes that connect directly (by one edge) to these remote homologues are collected as set $\{W_i\}$. Since homologous polypeptides have similar biological properties, we can characterize the probability of the query polypeptide i being in zone σ by that of its remote homologues: $P_i(\sigma) = \Sigma_j \delta(\sigma,\sigma'(w_{ij}))/\Sigma_j 1$, where $\sigma'(w_{ij}) = 0$ if a member w_{ij} of set $\{W_i\}$ belongs to helix-donut zone and $\sigma'(w_{ij}) = 1$ if w_{ij} is in strand-arc zone, step function $\delta(x,y)$ equals 1 for $x = y$ and $\delta(x,y) = 0$ otherwise, σ is 0 and 1 for the helix-donut zone and the strand-arc zone respectively. In the aspect of physics, every query segment can belong to both of the two states, but in different probabilities. Native structure is only one of its choices. A native state can jump to the other state, the pathogenic one with certain interchange probability. According to the secondary structure of a query protein, we can identify the native state $\sigma'(i)$ of a query segment i with the method described in Section 3.1 of reference [3]. The interchange probability can be evaluated as $Q_i = (1 - P_{i-1}(\sigma'(i)))P_i(\overline{\sigma'(i)})(1 - P_{i+1}(\sigma'(i)))$, where $\overline{\sigma'(i)} = 1 - \delta(\sigma'(i), 1)$ represent to the pathogenic state. Q_i is set to 0 if both helix and β-sheet exist in the residues of segment set $\{i-4, i-3, i-2, i-1, i, i+1, i+2, i+3, i+4\}$. If there are both helix and β-sheet residues in an enlarged segment, the two types of secondary structures could be interchangeable around segment i under normal thermal motion. Such facile interchange of the two state should not cause disease, and should be filtered. Then the polypeptides with a high value of Q_i are predicted to be switch region.

In our analysis, each residue is covered by at most 15 successive segments. To evaluate the significance of each residue site, we scored the interchange probability per site using the maximum interchange probability for the corresponding 15 polypeptides. Residues with the highest interchange probabilities, and their nearby residues that have interchange probabilities higher than 1.25 times of the overall probability at all position, are predicted to be switch region of corresponding protein.

3. Results

The clinical reports of actual cases of conformational diseases that are confirmed by definite nosogenesis are very scarce. Refs. [1,4–7] describe a total of 31 proteins responsible for various conformational diseases. Twenty-two of these have usable structural information. Since our method is based on knowledge of non-membrane proteins, this restricts the scope of the application, and five membrane or membrane-associated proteins are unsuitable for our method (amyloid-β precursor protein, cystic fibrosis transmembrane conductance, α-ketoacid dehydrogenase complex, β-hexosaminidase, α-synuclein). In Ref. [3], prion has ever been investigated to illustrate feasibility of our scheme. Here we analyzed all other classical proteins that are believed to be responsible for different conformational diseases, and identified regions that cover significant sites for pathogenic structural changes. Switch regions predicted are shown in red in Figs. 2–15, and marked by ↔ at corresponding probability cutoff. The conformational diseases involved in our test can be classified as blood disorders, cardiovascular disease, cerebrovascular disease, neuropathy, encephalopathy, cancer, blindness, and kidney disease, among others. For fibrinogen and fibrillin-1, we could not find any clinical reports regarding the residues involved in the PDB database. It was difficult to evaluate the predictions for these proteins and hence the results are not reported.

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

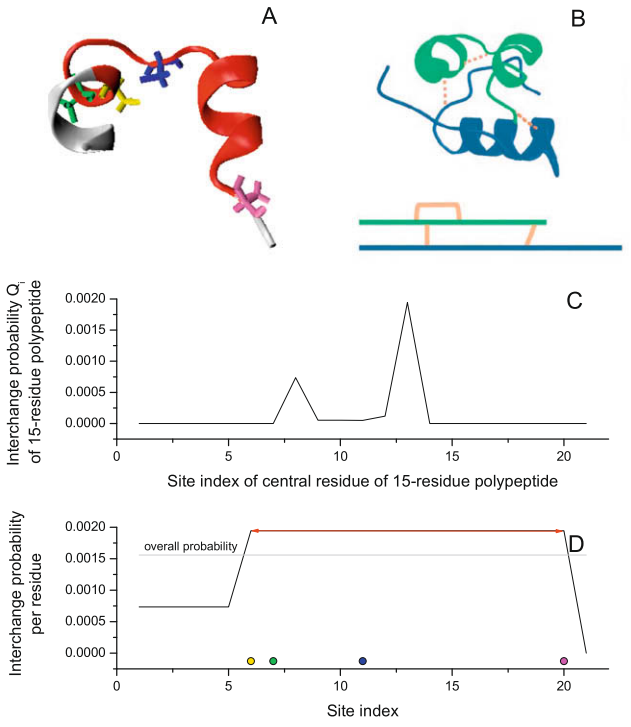

Fig. 2. Results of insulin (PDBID: 1ai0_A, 21 residues in length). (A) Structure of insulin. Cysteines responsible for disulfide bonds conserved in refolding are shown in bonds (6 yellow, 7 green, 11 blue and 20 magenta). (B) The three native disulfide bonds(gold) shown in insulin structure and topology diagram [9]. (C) Interchange probability for each 15-residue segment indexed by its central residue. (D) The interchange probability for each residue site. In (A) and (D), switch regions predicted are shown in red. This region involves every donor of disulfide bonds which act as constraints to refolding and contribute to initial aggregation of insulin. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

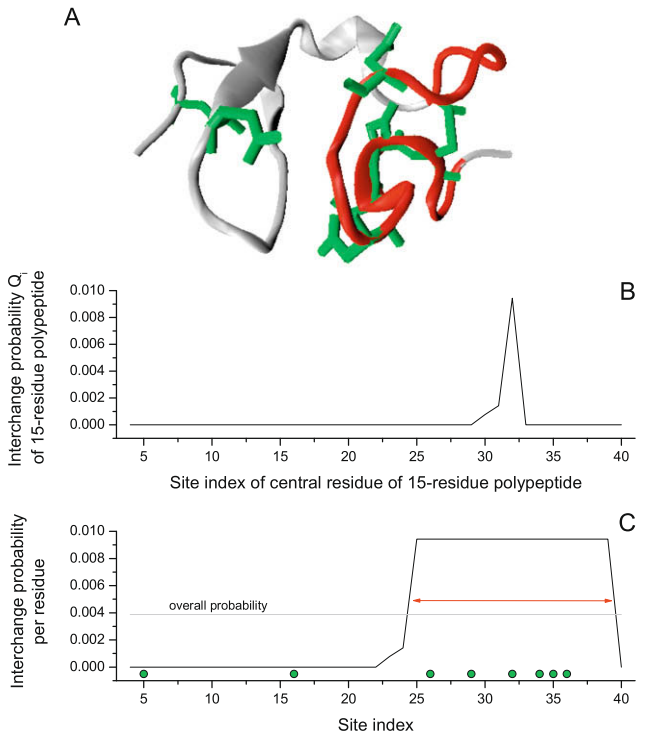

Fig. 3. Results of low-density lipoprotein receptor (PDBID: 1ajj_A, 37 residues in length). (A) Structure of low-density lipoprotein receptor. Residue sites related to pathogenic point mutations are shown in bonds, and colored in green. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. The region we predicted is a segment with highest density of disease-related sites, and should be switch region of low-density lipoprotein receptor. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

Initial sites of conformational changes should be the region where the switch role is evident. For some proteins, there are many residue sites for which disease-related mutations are clinically observed, i.e. where disease is induced. For a region dominating the initiation of pathogenic structural conversion, such disease-related site should be abundant. Consequently, it is rational to define a switch region according to density of disease-related site. Switch regions of low-density lipoprotein, apolipoprotein AI, superoxide dismutase, crystallins, and hemoglobin were identified in this way. As there are too many such sites, and differences in density are not obvious, only sites corresponding to highly unstable disease-related mutants were used in identifying switch sites of hemoglobin. For the other proteins, each switch region we defined is widely accepted in literatures.

3.1. Insulin

Insulin is a peptide hormone with extensive effects on metabolism and many body systems. Insulin injection is used medically to treat some forms of diabetes mellitus. Under solution conditions where the native state is destabilized, this largely helical polypeptide hormone can readily aggregate to form amyloid fibrils with a characteristic cross-β structure. Consequently, it is associated with a clinical syndrome, injection-localized amyloidosis [8].

It was revealed by mass spectrometry analysis that there is a character as insulin forming amyloid fibrils: The disulfide bonds of the native hormone are retained in the amyloid form, providing substantial constraints to refolding. Moreover according to the work of Jimènez et al., a segment donating such disulfide bond constraints plays a significant role in forming the initial aggregates of insulin amyloid fibrils, i.e. is a switch region [9]. As shown in Fig. 2, it coincides with our prediction that segment 6–20 is the switch region of insulin. Actually this segment is the minimum window that every cysteine responsible for the aforementioned disulfide bond constraints is involved, namely our prediction is very accurate.

3.2. Low-density lipoprotein (LDL) receptor

LDL receptor is a mosaic protein that mediates endocytosis of cholesterol-rich lipoprotein particles. The amino-terminal region of LDL receptor, which consists of seven tandemly repeated cysteine-rich modules (LDL-A modules), mediates binding to lipoproteins. Normally these LDL-A modules extend out into the extracellular fluid, seize the lipoproteins wherein. Then the lipoprotein arrested is imported into the cell by receptor-mediated endocytosis. Mutations of LDL receptor that affect this process result in failure to clear lipoprotein from the circulation, pathologically elevated blood cholesterol and premature heart disease.

Many point mutations that cause familial hypercholesterolaemia map to the fifth LDL-A module of the LDL receptor (LR5). As the module works far from membrane, largely not in a membrane-like environment, we can predict the switch region of LR5 with our method. As shown in Fig. 3, segment 25–39 is predicted as the switch region of LR5. This coincides with the observation that disease-related point mutations mainly map to a cluster of acidic residues near the carboxy-terminal end of LR5. There are totally eight residue sites for which disease-related point mutations

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

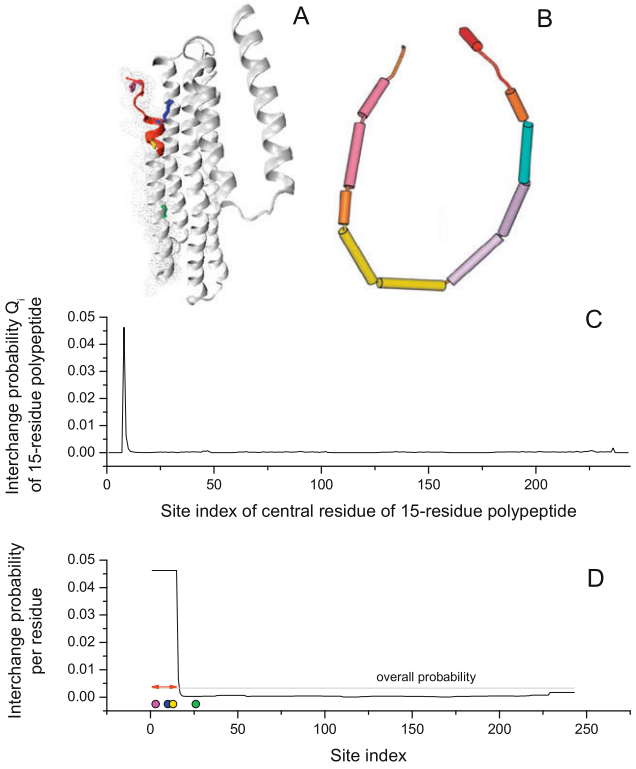

Fig. 4. Results of apolipoprotein AI (PDBID: 2a01_A, 243 residues in length). (A) Structure of lipid-free apolipoprotein AI. Helices A:1–43 are vital for structural stability of lipid-free Apo-AI, and are covered by dots. In helices A, disease-related sites are shown in bonds (3 magenta, 10 blue, 13 yellow and 26 green). (B) The elliptical loop formed by the △1–43 structure [11]. (C) Interchange probability for each 15-residue segment indexed by its central residue. (D) The interchange probability for each residue site. In (A) and (D), switch regions predicted are shown in red. Three out of four sites responsible for arising disease are in the region we predict. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

have been observed [10], i.e. illness is induced. Six of them are in the region we predicted.

3.3. Apolipoprotein AI

Apolipoprotein AI (Apo-AI) is the major protein component of high density lipoprotein (HDL) in plasma. The protein promotes cholesterol efflux from tissues to the liver for excretion. As an acceptor for sequential transfers of phospholipids, Apo-AI has two states in vivo (the lipid-free state and the lipid-bound one), with quite different conformations. The lipid-free Apo-AI is comprised of an N-terminal four-helix bundle and two C-terminal helices. Some mutations of Apo-AI can result in hereditary amyloidosis, such as familial amyloid polyneuropathy and familial visceral amyloid [1]. Since the amyloidosis is a consequence of protein aggregation, the lipid-free Apo-AI should be responsible for these diseases. Therefore, it is in the scope that we can cope with.

Stability of lipid-free Apo-AI should be vital for defending amylogenesis of Apo-AI. It is reported that mutations found in human amyloid deposits appear to occur more frequently at the amino terminus of Apo-AI. This is due to the fact that the N-terminal four-helix bundle, especially the helices A:1–43 are essential for the structural stability of lipid-free Apo-AI [11]. Structure of truncated protein △1–43 is quite different from wild-type fold. In consequent, there should be some residues in helices A governing the stability of native fold of lipid-free Apo-AI. In helices A, some point mutations at sites 3,10, 13, and 26 result in various clinical

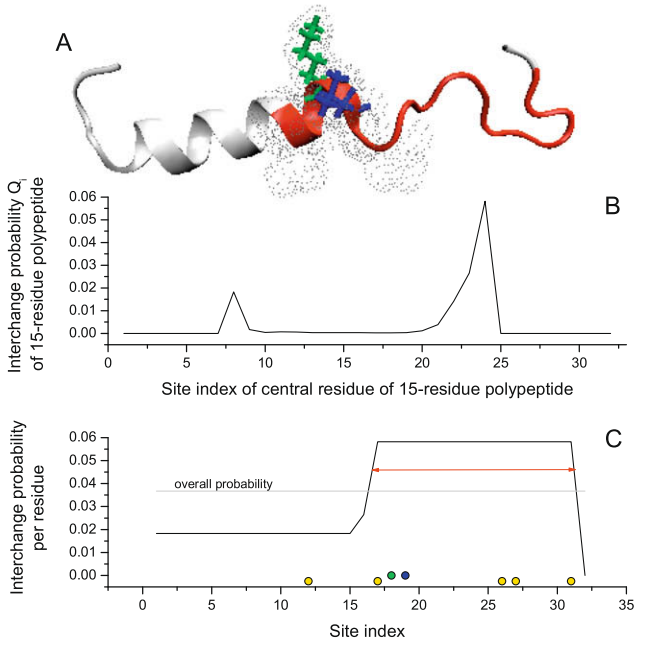

Fig. 5. Results of calcitonin (PDBID: 1byv_A, 32 residues in length). (A) Structure of calcitonin. Segment 15–21, shown in dots, governs fibril forming and bio-chemical properties of calcitonin. Wherein, the crucial residues 18 and 19 are shown in bonds, and colored in green and blue respectively. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. Sites related to the non-amyloidogenic analogue of human calcitonin, namely vital sites in inhibiting the pathogenic refolding are marked in yellow. In (A) and (C), switch regions predicted are shown in red. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

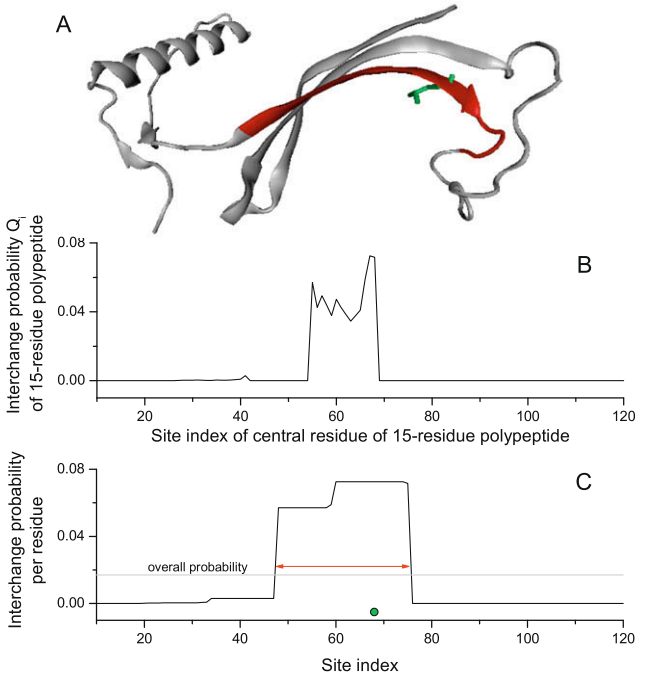

Fig. 6. Results of cystatin C (PDBID: 1g96_A, 111 residues in length). (A) Structure of cystatin C. In clinic reports, the most important site is at 68, shown in bonds and green. Mutation L68Q is associated with a severe conformational disease and causes massive amyloidosis, cerebral haemorrhage and death in young adults. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

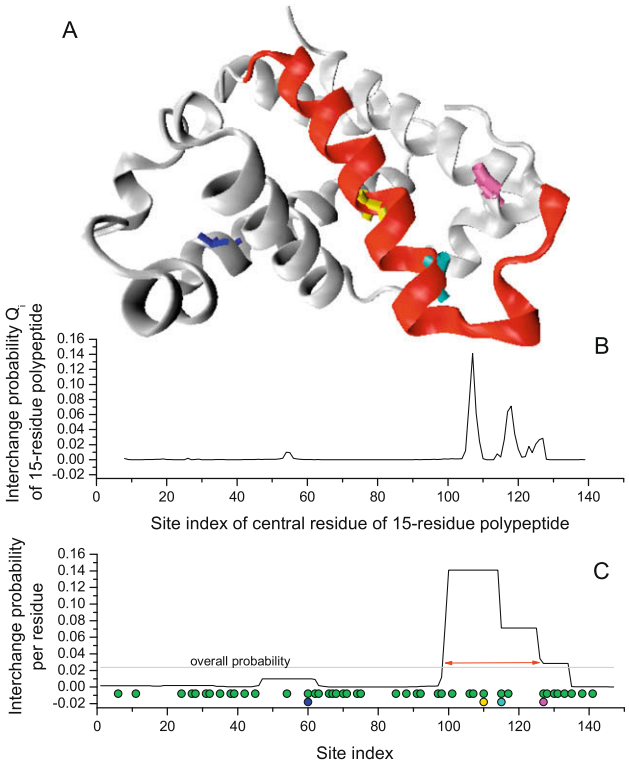

Fig. 7. Results of haemoglobin (PDBID: 1xz2_B, 146 residues in length). (A) Structure of haemoglobin β-chain. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. The 43 residue sites for which hemolytic anemia related point mutations have been observed are marked in green. Wherein the highly unstable mutants related sites are shown in various colors (60 blue, 110 yellow, 115 cyan, 127 magenta). In (A) and (C), switch regions predicted are shown in red. The highly unstable mutations are prone to occur in or close to the region we predicted. Therefore, disease-related stability of hemoglobin β-chain should be highly sensitive to the region we predicted. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

consequences [12]. These sites should be in switches in destabilizing native fold of lipid-free Apo-AI. As shown in Fig. 4, the switch region predicted, #1–15 is an inbuilt segment of helices A. Three out of four aforementioned disease-related sites are in the predicted region. It means our result is correct.

3.4. Calcitonin

Calcitonin is a 32-residue peptide hormone that is being produced by the C-cells of the thyroid and is mainly known for its hypocalcemic effects and the inhibition of bone resorption. Amyloid fibrils of human calcitonin were found to be associated with medullary carcinoma of the thyroid. Calcitonin has little secondary structure at room temperature. However, with a conformational conversion, calcitonin fibrils were found to be highly ordered, consisting of both helix and strand elements.

Recent work indicates a critical role of residue 15–21 for fibril forming and bioactivity of calcitonin [13,14]. In particular, the conformation and the topological features of side chains of residue 18 and 19 are strongly associated with the self-assembly state, binding affinity and the in vivo hypocalcemic potency of human calcitonin. Another interrelated report is that joint mutations: Y12L, N17H, A26N, I27T, A31T hamper the pathogenic refolding, and result in a non-amyloidogenic analogue of human calcitonin [15]. All the aforementioned sites are important for the disease-related stability of calcitonin. As shown in Fig. 5, segment 16–31 is identified

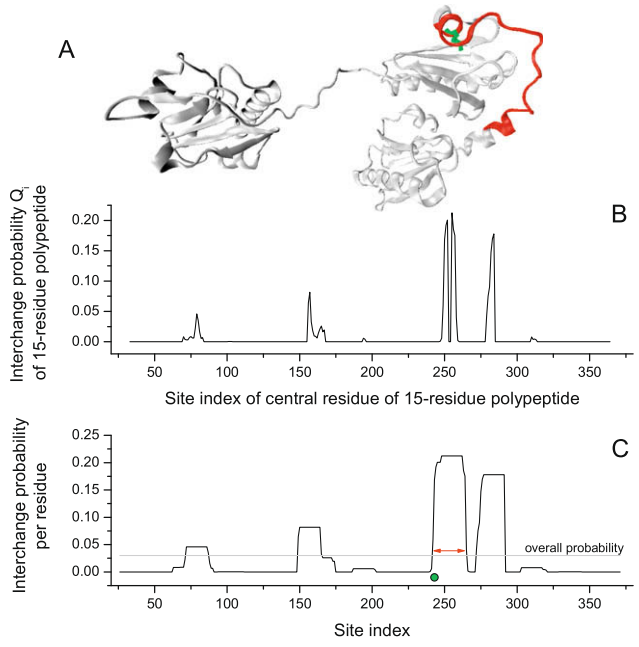

Fig. 8. Results of gelsolin (PDBID: 1rgi_A, 346 residues in length). (A) Structure of gelsolin. N-terminus or C-terminus of residue 243 is the cleavage site of gelsolin. The vital residue M243 in triggering amyloidogenesis is shown in bond and colored in green. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. Triggering site 243 is accurately predicted in gelsolin. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

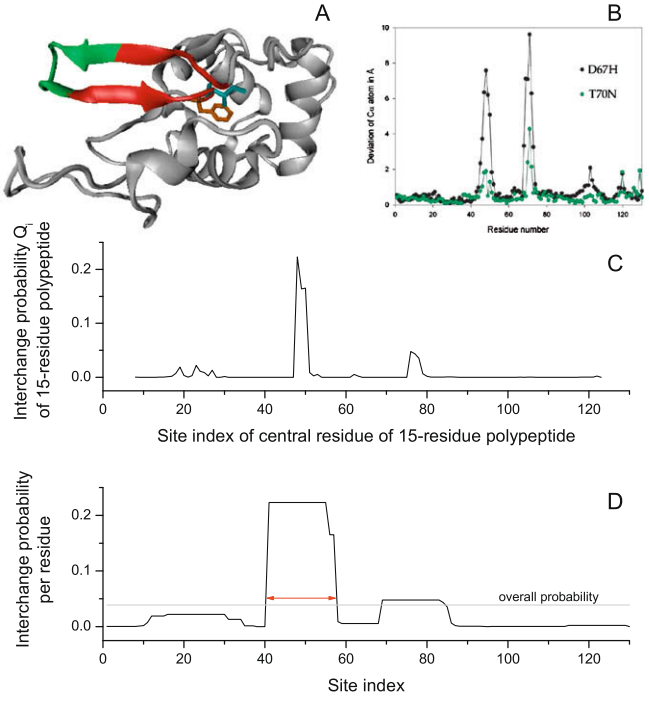

Fig. 9. Results of lysozyme (PDBID: 1w08_A, 130 residues in length). (A) Structure of lysozyme. According to Johnson et al., sites around 45–51, in green should be switch region of lysozyme. (B) A comparison of the crystallographic C^{α} atom deviations for T70N (green, PDB 1W08) and D67H (black, PDB 1LYY) variants, from their positions in the wild-type protein (PDB 1JSF) [22]. (C) Interchange probability for each 15-residue segment indexed by its central residue. (D) The interchange probability for each residue site. In (A) and (D), switch regions predicted are shown in red. Disease associated sites (56 cyan and 57 orange) involved in the switch region predicted are also shown in bonds. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

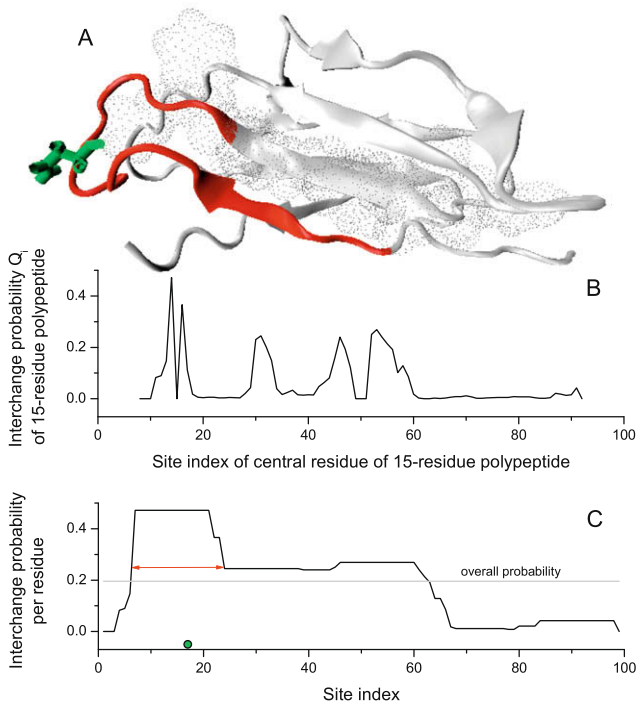

Fig. 10. Results of β_2 microglobulin (PDBID: 2vb5_A, 99 residues in length). (A) Structure of β_2 microglobulin. Amyloidogenic core fragment of β_2 microglobulin, 21–31 is shown in dots. The vital residue N17 is shown in bond and colored in green. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. There are three overlapped residues between the switch region predicted and the amyloid core. Residue N17 is important for the pathological mechanism of amyloid formation of wild-type β_2 microglobulin. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

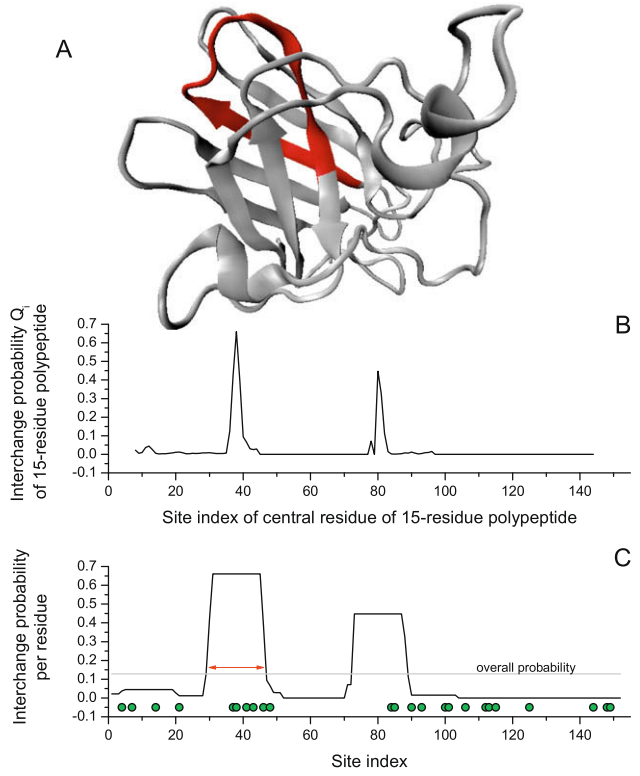

Fig. 11. Results of Cu/Zn type superoxide dismutase (PDBID: 2c9v_A, 153 residues in length). (A) Structure of Cu/Zn SOD enzyme. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. Disease-related sites are marked in (C), and in green. The density in segment 37–48 is at least 150% of those of other regions. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

as switch region in our prediction. This coincide with the aforementioned researches very well.

3.5. Cystatin C

Wild-type human cystatin C is a high-affinity inhibitor of some human cysteine proteases that belong to the papain family, such as cathepsins B, H, K, L, and S. In pathological processes, it forms part of the amyloid deposits in brain arteries that lead to cerebral angiopathy. Patients usually die in their teens from cerebral hemorrhage. The formation of amyloid cystatin C is claimed to be due to conformational changes in the monomer and subsequent domain swapping in the β-fibril structure [16].

According to clinical reports on cystatin C, the most important mutation is at 68 Leu → Gln, which is associated with a severe conformational disease and causes massive amyloidosis, cerebral hemorrhage and death in young adults [17,18]. Our analysis showed that a peak in interchange probability occurs at window 67 (Fig. 6). This means that residues 51–74 around site 68 are critical in this conformational disease, and are related to the initiation of structural changes in view of the double-zone feature of the polypeptide phase space.

3.6. Hemoglobin

Hemolytic anemia is a disorder in which destruction of red blood cells is faster than their production by bone marrow. Many cases of this disease are believed to be due to the presence of unstable hemoglobin that can change its structure and result in disorder. There are various causes of hemoglobin instability, such as single-point mutations, insertion or deletion of amino-acids, frame shifts, etc. Unstable hemoglobin molecules lead to varying degrees of hemolysis [19].

Here we analysed the β-chain of human hemoglobin that is believed to be responsible for several hemolytic anemia. The detailed results are shown in Fig. 7. Two successive peaks for the interchange probability occur at windows 107 and 118. The two polypeptides overlap and have much higher interchange probabilities (at least two-fold) than the other sites. This means that the switch region for hemoglobin is greater than 15 residues and should be extended. Thus, residues 99–125 were predicted to be prone to conformational changes leading to disorder. This result coincides with clinical reports. According to the database in reference [20], there are totally 43 residue sites for which hemolytic anemia related point mutations have been observed. These sites are interspersed along the 146-residue sequence. In all the corresponding variants, there are four highly unstable mutants (V60E, L110P, A115D, Q127R). Three out of the four are in or close to the region we predicted. It means that the switch region predicted is significant for the structural stability of hemoglobin β-chain.

3.7. Gelsolin

Gelsolin is an actin-binding protein that is a key regulator of actin filament assembly and disassembly. Wild-type gelsolin is not associated with any amyloid pathology; however, inheritance of some types of mutations, e.g. D187N or D187Y, confers 100% penetrance of Finnish hereditary systemic amyloidosis which is characterized by extensive skin, arterial, neurologic, and ophthalmologic amyloid

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

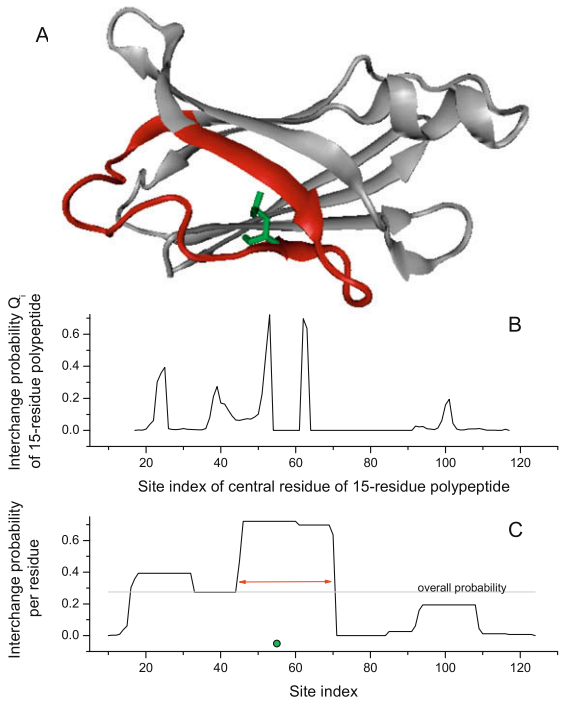

Fig. 12. Results of transthyretin (PDBID: 1dvq_A, 115 residues in length). (A) Structure of transthyretin. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. Site 55 in green is the most important site according to clinic reports. Mutation 55 Leu → Pro is the most notorious mutant, and cause early-onset familial amyloidotic polyneuropathy with onset of clinical disease appearing approximately 20 years of age. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

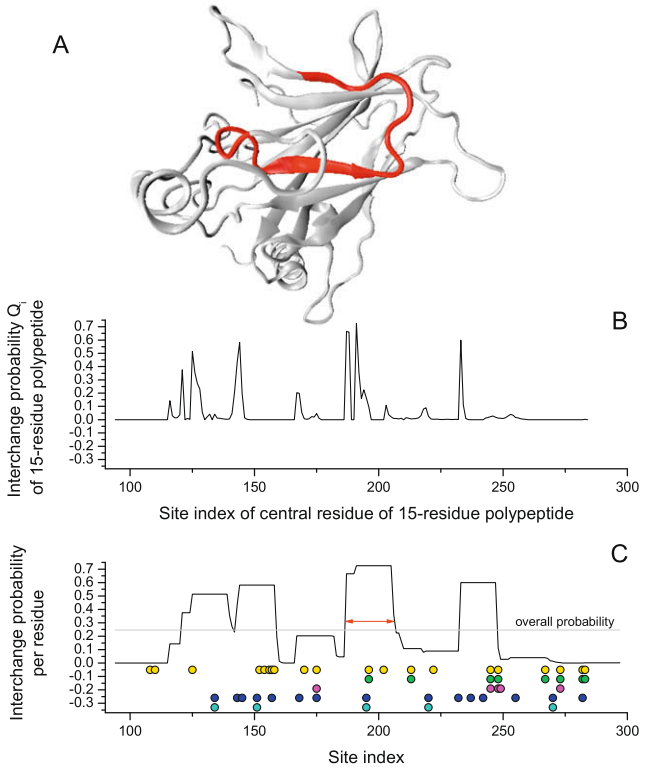

Fig. 13. Results of tumor suppressor protein p53 (PDBID: 2fej_A, 204 residues in length). (A) Structure of p53. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. We found p53 is extremely complicated: ⟨A mass of sites are sensitive to p53's biological activity⟩ The twenty sites for which the activity of mutants decreases more than half of that of wild-type protein are marked in yellow. Some variants of zero biological activity are produced by point mutation at sites marked in green. ⟨Nearly every site corresponds to some disease-related point mutations respectively⟩ Magenta points show sites of the top six most frequently mutated residues in human cancer [38]. ⟨Various regions are significant for the disease-related stability of p53⟩ Blue points are sites corresponding to the highly destabilizing mutants reported in [40]. Positions related to the top five most highly destabilizing variants are marked in cyan. In (A) and (C), switch regions predicted are shown in red. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

deposition. Unlike some disease-related proteins which undergo amyloidogenic conformational changes in native fold, gelsolin is associated with amyloidosis due to aberrant cleavage of its precursor protein, i.e. the formation of 70- and 71-residue amyloidogenic gelsolin fragment found in patients. Abnormal cleavage is a direct causation of the disease.

As shown in Fig. 8, segment 243–264 is predicted to be switch region of gelsolin. It means that disease-related mutation should confer region 243–264 a notable conformational change which produce a sufficient condition for aberrant cleavage, e.g. peptide unique to hydrolase digest. This coincides with the discovery of Page et al. that gelsolin amyloidogenesis is triggered by metalloendoprotease cleavage [21]. The cleavage site is fitly either A^{242}–M^{243} or M^{243}–L^{244}.

3.8. Lysozyme

Lysozyme is an antibacterial protein for which mutations are associated with familial visceral amyloidosis in the liver, spleen, kidneys, and other internal organs. There are five known mutations in the human lysozyme gene that give rise to six variant proteins, 56 Ile → Thr, 57 Phe → Ile, 64 Trp → Arg, 67 Asp → His, 70 Thr → Asn, and the double mutation F57I& T70N. All the variants apart from T70N have been detected in association with amyloid deposits in various human patients.

There are two candidate switch regions, around sites 57 and 67. As shown in Fig. 9B, compared with structure of wild-type protein, mutation T70N can result in considerable structural rearrangement at sites 68–75 [22] without causing conformational disease. Therefore, the second site is likely not important for the initiation of pathogenic structural changes and can be excluded. In fact, according Fig. 9B, there is some structural rearrangement at sites 45–51 in the T70N variant, but this is not large enough to cause conformational disease. However the structural rearrangement in this region is very strong for the disease-related mutant D67H, the amyloid donor. This indicates that sites 45–51 correspond to a switch region for lysozyme. As shown in Fig. 9, successive peaks for the interchange probability occur at windows 48–50, so that polypeptide 41–57 should be significant in the initiation of structural changes, in agreement with previous research.

3.9. β_2 microglobulin

β_2 microglobulin is a non-polymorphic light chain of the class I major histocompatibility complex (MHC-I) that plays an important role in the immune system, autoimmunity, and reproductive success [23]. As part of its normal catabolic cycle, β_2 microglobulin dissociates from MHC-I and is transported in serum to the kidneys, where the majority of the protein is degraded. If there is renal failure whereby β_2 microglobulin does not pass through the dialysis membrane, then its clearance from serum is disrupted. This results in an increase in β_2 microglobulin. When a high blood level is maintained for more than 10 years, the protein then self-associates

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

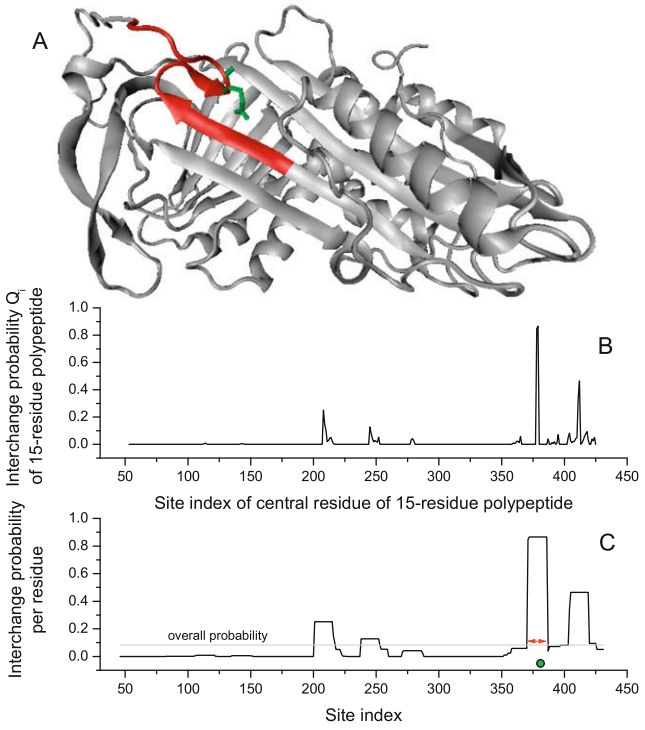

Fig. 14. Results of serpins (PDBID: 1e04_A, 386 residues in length). (A) Structure of serpins. (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. Site 381, in green plays an important role in stabilising native, inserted, and activated state of serpins. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

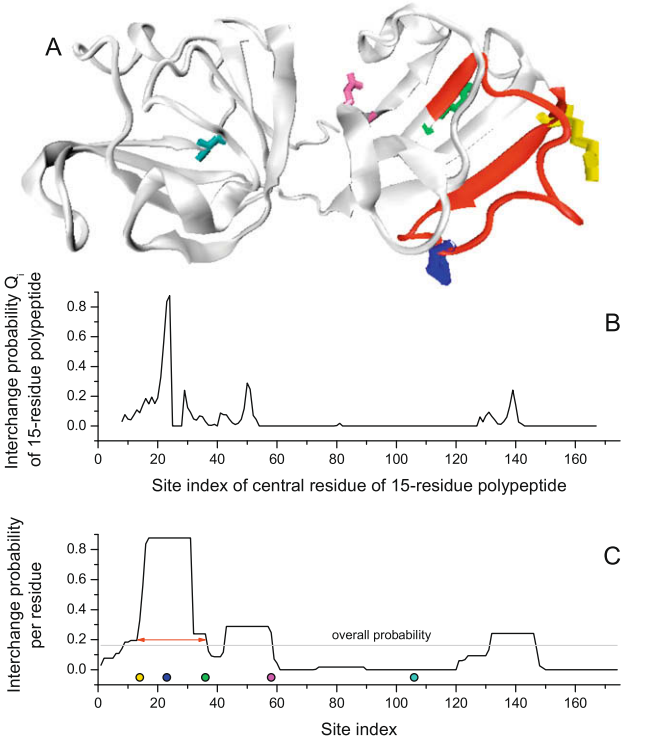

Fig. 15. Results of crystallin (pdbid:1hk0_X, 173 residues in length). (A) Structure of human gamma-D crystallin. The five disease-related sites are shown in bonds (14 yellow, 23 blue, 36 green, 58 magenta, and 106 cyan). (B) Interchange probability for each 15-residue segment indexed by its central residue. (C) The interchange probability for each residue site. In (A) and (C), switch regions predicted are shown in red. (For interpretation of color mentioned in this figure the reader is referred to the web version of the article.)

to form amyloid fibrils, causing dialysis-related amyloidosis [24,25].

It is believed that the formation of amyloid fibrils of β_2 microglobulin accompanies a significant conformational change. There are two successive peaks for the interchange probability at central sites 14 and 16 in Fig. 10, indicating that sites 7–23 are likely involved in the initiation of structural changes in β_2 microglobulin. This should correlate with the observation that fragment 21–31 is the well-known amyloidogenic core fragment of β_2 microglobulin [26–28]. Different from some other proteins, the wild-type β_2 microglobulin can aggregate due to the influence of ageing. In such process, acidification, e.g. in site 17(N → D), is necessary to form amyloid fibrils from both wild-type β_2 microglobulin and its variants [29]. It means residues around 17 is significant in triggering pathogenic refolding for both wild-type microglobulin and its variants.

3.10. Superoxide dismutase, SOD

The enzyme superoxide dismutase is metalloprotein which catalyzes the dismutation of superoxide into oxygen and hydrogen peroxide. Therefore, it is an important antioxidant defense in nearly all cells exposed to oxygen. There are three major families of superoxide dismutase, depending on the metal cofactor: Cu/Zn type, Fe/Mn type, and Ni type. Some mutations in Cu/Zn SOD enzyme can cause familial amyotrophic lateral sclerosis.

According to reference [30], there are totally 26 residue sites for which disease-related point mutations have ever been reported. As shown in Fig. 11, segment 37–48 is the region with the highest density of the 26 residue sites. This coincides with our prediction that segment 30–46 should be switch region of pathogenic structural changes for Cu/Zn SOD enzyme.

3.11. Transthyretin

Transthyretin (TTR) is a serum and cerebrospinal fluid carrier of the thyroid hormone thyroxine. It also acts as a carrier of retinol (vitamin A) through an association with retinol binding protein. Amyloid deposition of TTR is associated with several diseases, such as senile systemic amyloidosis, familial amyloid neuropathy, and familial cardiac amyloid [1]. The fibrillar structure resulting from self-association of an abnormal conformation of TTR is thought to be the causative agent in these disorders. The majority of TTR-associated amyloidoses are due to single amino-acid substitutions. In senile systemic amyloidosis, the non-mutated protein is present in amyloid fibrils [31]. However, the mechanism that by which normally soluble TTR tetramers are converted into insoluble amyloid fibrils remains largely unknown.

Here we analyzed normal human TTR to identify sites involved in the initial structural changes in this protein. As shown in Fig. 12, polypeptide 46–69 (central residue corresponding to the two successive peaks at sites 53 and 62) was identified as the switch region for conformational changes in TTR. Clinical data demonstrated that the mutation 55 Leu → Pro can cause early-onset familial amyloidotic polyneuropathy [32]. According to clinical reports, L55P is the most notorious mutant, with onset of clinical disease appearing approximately 20 years of age. In comparison, the age of onset is approximately 30 years for V30M carriers and 80 years for wild-type subjects. Analysis based on the crystal structure of the L55P mutant showed that site 55 is important site in the pathway for TTR polymerization to amyloid fibrils [33]. Amyloidogenic regions experimentally determined to date are 10–19 [34] and 105–115 [35], but no segments covering site 55 have been identified so far. As shown in this example, amyloid-related

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

mutations are not necessarily involved in aggregation-prone regions. So it is still not clear whether switch sites occur in hot spots of aggregation [36,37] or not.

3.12. Tumor suppressor protein p53

p53 is a transcription factor in multicellular organisms, where it regulates the cell cycle and thus functions as a tumor suppressor in preventing cancer. As such, p53 has been described as "the guardian of the genome", referring to its role in conserving stability by preventing genome mutation. In about 50% of human cancers, p53 is inactivated as a result of missense mutation in the p53 gene.

Actually, p53 is the most complicated molecule we have ever coped with. As shown in Fig. 13, there are totally twenty sites for which the activity of mutants decreases more than 50% [38]. Such sites are interspersed along the 289-residue fold. All mutants without biological activity locate in the N-terminal half (site index >195) of p53 sequence. Moreover, five of the top six amino-acid residues that are most frequently mutated in human cancer are in the N-terminal half (Arg-175, Gly-245, Arg-248, Arg-249, Arg-273, and Arg-282). It means the N-terminal half is significant for conserving the activity of p53 (this is largely due to the property conservation of DNA-binding core domain [39]). We predicted region 187–205 to be switch region of p53. The peaks in the N-terminal half are higher than those of the other half. This fits with the aforementioned knowledge qualitatively. Whereas, features in the stability of p53 is extremely complicated. In Fig. 13, there are several peaks with similar standards of interchange probability. Each of them involves several sites for which the highly destabilizing mutants have been reported [40]. As such, there should be several switch regions in p53, interfering accuracy of our method by unsuitable basic hypothesis (we suppose there is only one switch region in protein).

3.13. Serpins

Serpins are a family of proteins that inhibit proteases via a profound conformational change that irreversibly locks the protease and serpin together. The normal physiological functions of serpins are based on such highly specific transitions in the natural conformation. Premature conversion of the protein structure will result in a deficiency or dysfunction of the inhibition of proteases, then fibril formation and quite different disease consequences such as emphysema, cirrhosis, and thromboembolic disease [1].

Results of our analysis were shown in Fig. 14. Initiation sites for serpins are predicted to be around site 379 (polypeptide 372–386), in agreement with the results of Johnson et al., who reported that wild-type residue Glu-381 plays an important role in stabilising the native, inserted, and activated states of serpin proteins [41]. Actually the predicted region is on the N-terminus of reactive loop of serpins which is vital for the biological properties of the protein.

3.14. Crystallin

Crystallin is a water-soluble structural protein found in the lens of the eye. It is the major protein of the eye lens, accounting for the transparency of lens. Mutations and aging of crystallins cause cataracts, the predominant cause of blindness in the world. In human gamma-D crystallin, there are totally five residue sites for which mutants associated with congenital cataracts have ever reported, i.e. R14, P23, R36, R58, and E106. Three of the five are in the segment 14–36 which we predicted as switch region of gamma-D crystallin (Fig. 15). Moreover, the threonine substitution in site 23 is reported as the cause of pivotal local conformational and dynamic differences in human gamma-D crystallin [42]. All these evidences prove our result is correct.

3.15. Summary of results

Here we identify the significant residue positions in an independent aspect, physics that would be helpful in optimizing the knowledge contributed by clinical reports. As shown in most of the above examples, the prediction results match clinical reports very well. Besides the neutral prediction for the extremely complicated p53, 14 out of 15 proteins (including prion) were correctly predicted, both sensitivity and specificity achieved 93% for proteins in body fluid. As we aim to identify the significant regions of proteins, provide a guide of in-depth investigation, the evaluation of accuracy at segment level is reasonable. There are a total of 2196 residues in the test set. Only 12 percentages residues (264) were predicted as residues in switch regions. Therefore, the algorithm is highly sensitive and specific.

Another method of evaluation is to estimate the statistical significance of the predictions for residues that are tightly associated with conformational diseases (RTACD). For a N residues protein, if the m residues of the switch region predicted are randomly selected, the probability of covering u out of the U RTACDs can be calculated as $\chi = C_U^u C_{N-U}^{m-u}/C_N^m$. As shown in Table 1, such probability is

Table 1
Statistical analysis of the predictions. U is the total number of residues that are tightly associated with conformational diseases (RTACD). u is the counts of RTACD that are in the region predicted.

Proteins	Sequence length, N	Length of switch region predicted, $m(\frac{m}{N})$	RTACD coverage of the unpredicted & predicted regions, $\frac{U-u}{N-m}$ & $\frac{u}{m}$		Prediction coverage of RTACD, $\frac{u}{U}$	Probability of the prediction estimated by the coverage of RTACD, χ	Coverage of the unproved pre dictions, $\frac{m-u}{m}$
Insulin	21	15 (0.71)	0	0.27	4/4 = 1)	0.23	0.73
LDL receptor	37	15 (0.41)	0.091	0.4	6/8 = 0.75	0.03	0.6
Apo-AI	243	15 (0.06)	0.0044	0.2	3/4 = 0.75	0.00073	0.8
Calcitonin	32	16 (0.50)	0	0.13	2/2 = 1	0.25	0.875
Cystatin C	111	24 (0.22)	0	0.042	1/1 = 1	0.22	0.96
Hemoglobin	147	27 (0.18)	0.0083	0.11	3/4 = 0.75	0.018	0.89
Gelsolin	346	22 (0.06)	0	0.045	1/1 = 1	0.064	0.95
Lysozyme	130	17 (0.13)	0.018	0.12	2/4 = 0.5	0.074	0.88
$\beta_2$microglobulin	99	17 (0.17)	0	0.059	1/1 = 1	0.17	0.94
SOD	152	15 (0.10)	0.15	0.4	6/26 = 0.23	0.018	0.6
Transthyretin	115	24 (0.21)	0.011	0.042	1/2 = 0.5	0.33	0.96
p53	204	19 (0.09)	0.022	0.053	1/5 = 0.2	0.33	0.95
Serpins	386	15 (0.04)	0	0.067	1/1 = 1	0.038	0.93
Crystallin	173	23 (0.13)	0.013	0.13	3/5 = 0.6	0.016	0.87
	Sum	Sum (mean ± sd)	Mean ± sd	Mean ± sd	Mean ± sd	Mean ± sd	Mean ± sd
	2196	264 (0.22 ± 0.19)	0.022 ± 0.043	0.15 ± 0.13	0.73 ± 0.29	0. 13 ± 0.12	0.85 ± 0.13

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

about 13 percentages. It means that the present algorithm can identify the significant signal in approximately eight times of the ability of random dicing.

Although per residue prediction is beyond the scope of the present work, we found that the RTACDs are abundant in the region predicted. The coverage of RTACDs is about 73% in our prediction, that is, most RTACDs are covered by the region predicted, where about 15% residues are RTACDs. But in the other regions, the ratio decreases into 2.2% approximately. There are about 85% sites predicted, but unproved for their vital role. It provides further opportunity of in-depth research.

In an aspect of evolution, the interchange probability evaluates incidence of a disease under selection pressure. As shown in Fig. 16, a graph of the interchange probability of switch regions for different proteins reveals that cases with low interchange probabilities are highly fatal (expect injection-localized amyloidosis induced artificially). For prion, cystatin C and hemoglobin, without clinical treatment, patients die at a very young age and usually have no opportunity to transmit their genes to offspring. As abundantly expressed in vital tissues, these proteins must be very stable. Once misfolding occurs the result is fatal. Protein evolution ensures such stability by low interchange probability.

4. Discussions

In this work, we present a prediction scheme that points out the most important sites for the source of pathogenic structural change. Such original place could provide ideal target for clinical therapy, serve as probe binding site for high sensitive detect, and etc. Moreover, it can evaluate the incidence, i.e. risk level of disease arisen by a protein from an evolutionary point of view. Due to its ability in clarifying target sites, filtering secondary factors, decreasing the scope of in-depth study, this algorithm would aid CD research drastically.

Protein aggregation which often results in tissue deposition is one of the well-known pathogenesis of conformational disease. It is a consequence of the initial misfolding at switch sites. To investigate mechanism of the formation of amyloid, many efforts were made. It was found that some sequences are much more amyloidogenic than others, and there are aggregation-prone regions, "hot spots" of fibril-forming, which are considered to be responsible for aggregation [36,37]. These data has inspired a number of algorithms and models in the prediction of aggregation propensities of protein [43–52]. While, it is still obscure whether switch sites lie in hot spots of aggregation or not. As shown in the example of transthyretin, amyloid-related mutations are not necessarily involved in aggregation-prone regions. Compared with the small counts of switch sites, aggregation-prone sites are abundant in disease-related proteins. Consequently, according to de Groot et al. [47], at least one-third of residues are involved in predicted hot spots. Therefore, switch site is more specific. To the best of our knowledge, the present work is the first algorithm that can predict switch regions for various conformational diseases.

As shown by the above results, this method is very suitable for proteins in body fluid. However, for membrane and membrane-associated protein, the predictions are less satisfactory. There are two possible reasons. Firstly, our method is based on knowledge of non-membrane proteins and may thus be unsuitable for solving problems related to membrane. Secondly, the membrane environment is more complicated than that of body fluid for proteins. Interactional effects between proteins and the membrane environment complicate the problem. To overcome this issue, our future research will focus on knowledge based on membrane systems.

Many proteins are caused by structural change. Besides of those identified by clinical analysis, there are also lots of unknown cases. As a prerequisite for in-depth research, clinical information is very important. However, as sophisticated techniques and trained researchers are necessary for experiments in identifying disease-related proteins, clinical information is not sufficient and is difficult to obtain. As thus a systematic study of conformational diseases is beyond the scope of previous technical approaches. While, in combination with medical research, our work could have a profound impact to such situation. For example, a mutation occurs in switch region means a potential causation of conformation disease. Therefore, the algorithm enlarges the research scope by identifying more highly suspect but unknown disease-related proteins and diminishes the residue counts for in-depth investigation, namely improves CD research in strategic and in tactical simultaneously. Health of human being will potentially be boosted. Due

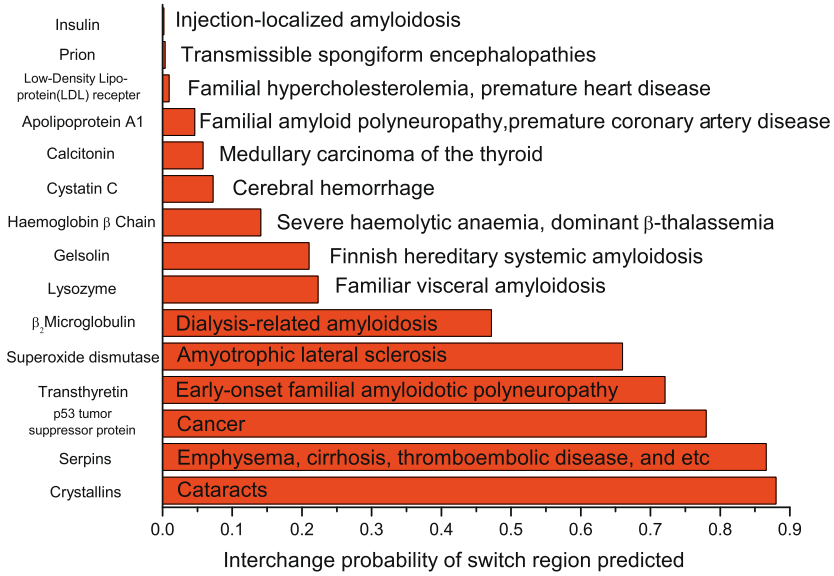

Fig. 16. Interchange probability for the switch region predicted for different proteins. Most cases with low interchange probability are highly fatal. Since the interchange probability can approximately characterize the incidence of a disease from an evolutionary point of view, this means that fatal cases are filtered due to the selection pressure of evolution.

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003

to the high success rate, present algorithm is a powerful tool in coping with some urgent and knowledge lack cases in human health, such as the highly pathogenic avian (H5N1) flu and 2009 A (H1N1) influenza [53].

In modern bioscience, scientists are compelled to deal with a mass of candidate molecules. There is hot demand for efficient method to help them discarding unnecessary studies, and grasping something essential. This method fits such desirability, and would attract many users. Since key residues are pointed out with the algorithm, it also provides useful entry in performing interdisciplinary researches for the scientists outside of the field.

To ensure a healthy development of modern biology, a patent is applied for corresponding method. We encourage pure scientific research. Contact authors when the method is to be used.

We are grateful to Professor Yuan-Kai Hong and Dr. Ming Li for their assistance and recommendations. This work was jointly supported by the National High-tech R&D Program of China (863 Program, Grant No. 2007AA021803), National Basic Research Program of China (973 Program, Grant No. 2007CB310500), and National Natural Science Foundation of China, No. 10704077.

References

[1] Carrell RW, Lomas DA. Conformational disease. Lancet 1997;350:134–8.
[2] Dobson CM. Protein misfolding, evolution and disease. Trends Biochem Sci 1999;24:329–32.
[3] Xin L, Ya-Pu Z. Donut-shaped fingerprint in homologous polypeptide relationships—a topological feature related to pathogenic structural conversion of conformational disease. J Theor Biol 2009;258:294–301.
[4] Carrell RW, Gooptu B. Conformational changes and diseases-serpins, prions, and Alzheimer's. Curr Opin Struct Biol 1998;8:799–809.
[5] Soto C. Protein misfolding and disease; protein refolding and therapy. FEBS Lett 2001;498:204–7.
[6] Kelly JW. Alternative conformations of amyloidogenic proteins govern their behavior. Curr Opin Struct Biol 1996;6:11–7.
[7] Thomas PJ, Qu BH, Pedersen PL. Defective protein folding as a basis of human disease. Trends Biochem Sci 1995;20:456–9.
[8] Dische FE, Wernstedt C, Westermark GT, Westermark P, Pepys MB, Rennie JA, et al. Insulin as an amyloid-fibril protein at sites of repeated insulin injections in a diabetic patient. Diabetologia 1988;31:158–61.
[9] Jiménez JL, Nettleton EJ, Bouchard M, Robinson CV, Dobson CM, Saibil HR. The protofilament structure of insulin amyloid fibrils. Proc Natl Acad Sci USA 2002;99:9196–201.
[10] Fass D, Blacklow S, Kim PS, Berger JM. Molecular basis of familial hypercholesterolaemia from structure of LDL receptor module. Nature 1997;388:691–3.
[11] Ajees AA, Anantharamaiah GM, Mishra VK, Hussain MM, Murthy HMK. Crystal structure of human apolipoprotein A-I: insights into its protective effect against cardiovascular diseases. Proc Natl Acad Sci USA 2006;103:2126–31.
[12] Frank PG, Marcel YL. Apoliproprotein A-I: structure–function relationship. J Lipid Res 2000;41:853–72.
[13] Haspel N, Zanuy D, Ma B, Wolfson H, Nussinov R. A comparative study of amyloid fibril formation by residues 15–19 of the human calcitonin hormone: a single β-sheet model with a small hydrophobic core. J Mol Biol 2005;345:1213–27.
[14] Kazantzis A, Waldner M, Taylor JW, Kapurniotu A. Conformationally constrained human calcitonin (hCt) analogues reveal a critical role of sequence 17–21 for the oligomerization state and bioactivity of hCt. Eur J Biochem 2002;269:780–91.
[15] Andreotti G, Vitale RM, Avidan-Shpalter C, Amodeo P, Gazit E, Motta A. Designing aggregation-resistant bioactive peptides by three-dimensional structure homology with a non-amyloidogenic analogue. Submitted for publication.
[16] Janowski R et al. Human cystatin C, an amyloidogenic protein, dimerizes through threedimensional domain swapping. Nat Struct Biol 2001;8:316–20.
[17] Abrahamson M. Molecular basis for amyloidosis related to hereditary brain hemorrhage. Scand J Clin Lab Invest 1996;226:47–56.
[18] Olafsson I, Grubb A. Hereditary cystatin C amyloid angiopathy. Amyloid Int J Exp Clin Invest 2000;7:70–9.
[19] Outeirino J, Casey R, White JM, Lehmann H. Haemoglobin Madrid beta 115 (G17) alanine–proline: an unstable variant associated with haemolytic anaemia. Acta Haematol 1974;52:53–60.
[20] Available from: http://globin.cse.psu.edu/html/huisman/variants/beta/index.html.
[21] Page LJ, Suk JY, Huff ME, Lim HJ, Venable III J, Kelly JY, et al. Metalloendoprotease cleavage triggers gelsolin amyloidogenesis. EMBO J 2005;24:4124–32.
[22] Johnson RJK et al. Rationalising lysozyme amyloidosis: insights from the structure and solution dynamics of T70N lysozyme. J Mol Biol 2005;352:823–36.
[23] Bjorkman PJ et al. Structure of the human class I histocompatibility antigen, HLA-A2. Nature 1987;329:506–12.
[24] Floege J, Ketteler M. β2-Microglobulin-derived amyloidosis: an update. Kidney Int 2001;59:164–71.
[25] Gejyo F et al. A new form of amyloid protein associated with chronic hemodialysis was identified as beta 2-microglobulin. Biochem Biophys Res Commun 1985;129:701–6.
[26] Hasegawa K et al. Amyloidogenic synthetic peptides of β2-microglobulina role of the disulfide bond. Biochem Biophys Res Commun 2003;304:101–6.
[27] Hiramatsu H, Goto Y, Naiki H, Kitagawa T. Core structure of amyloid fibril proposed from IR-microscope linear dichroism. J Am Chem Soc 2004;126:3008–9.
[28] Hiramatsu H, Goto Y, Naiki H, Kitagawa T. Structural model of the amyloid fibril formed by β_2-microglobulin #21–31 fragment based on vibrational spectroscopy. J Am Chem Soc 2005;127:7988–9.
[29] Kad NM, Thomson NH, Smith DP, Smith DA, Radford SE. β_2-microglobulin and its deamidated variant, N17D form amyloid fibrils with a range of morphologies in vitro. J Mol Biol 2001;313:559–71.
[30] de Belleroche J, Orrell R, King A. Familial amyotrophic lateral sclerosis/motor neurone disease (FALS): a review of current developments. J Med Genet 1995;32:841–7.
[31] Saraiva MJM. Transthyretin mutations in health and disease. Hum Mutat 1995;5:191–6.
[32] Jacobson DR, McFarlin DE, Kane I, Buxbaum JN. Transthyretin Pro55, a variant associated with early-onset, aggressive, diffuse amyloidosis with cardiac and neurologic involvement. Hum Genet 1992;89:353–6.
[33] Sebastiao MP, Saraiva MJ, Damas AM. The crystal structure of amyloidogenic Leu55 → Pro transthyretin variant reveals a possible pathway for transthyretin polymerization into amyloid fibrils. J Biol Chem 1998;273:24715–22.
[34] Jarvis JA, Kirkpatrick A, Craik DJ. 1H NMR analysis of fibril-forming peptide fragments of transthyretin. Int J Pept Protein Res 1994;44:388–98.
[35] Jaroniec CP, MacPhee CE, Astrof NS, Dobson CM, Griffin RG. Molecular conformation of a peptide fragment of transthyretin in an amyloid fibril. Proc Natl Acad Sci USA 2002;99:16748–53.
[36] Ventura S et al. Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. Proc Natl Acad Sci USA 2004;101:7258–63.
[37] Ivanova MI, Sawaya MR, Gingery M, Attinger A, Eisenberg D. An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. Proc Natl Acad Sci USA 2004;101:10584–9.
[38] Available from: http://p53.free.fr/Database/p53_database.html.
[39] Joerger AC, Ang HC, Fersht AR. Structural basis for understanding oncogenic p53 mutations and designing rescue drugs. Proc Natl Acad Sci USA 2006;103:15056–61.
[40] Available from: http://www-p53.iarc.fr/stability.html.
[41] Johnson DJD, Huntington JA. The influence of hinge region residue Glu-381 on antithrombin allostery and metastability. J Biol Chem 2004;279:4913–21.
[42] Jung J, Byeon IJ, Wang Y, King J, Gronenborn A. The structure of the cataract causing P23T mutant of HgD crystallin exhibits local distinctive conformational and dynamic changes. Biochemistry. doi: 10.1021/bi802292q.
[43] Lopez de la Paz M, Serrano L. Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci USA 2004;101:87–92.
[44] Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 2004;22:1302–6.
[45] Yoon S, Welsh WJ. Detecting hidden sequence propensity for amyloid fibril formation. Protein Sci 2004;13:2149–60.
[46] Pawar AP, Dubay KF, Zurdo J, Chiti F, Vendruscolo M, Dobson CM. Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J Mol Biol 2005;350:379–92.
[47] de Groot NS, Pallarés I, Avilés FX, Vendrell J, Ventura S. Prediction of "hot spots" of aggregation in disease-linked polypeptides. BMC Struct Biol 2005;5:18.
[48] Bemporad F, Calloni G, Campioni S, Plakoutsi G, Taddei N, Chiti F. Sequence and structural determinants of amyloid fibril formation. Acc Chem Res 2006;39:620–7.
[49] Caflisch A. Computational models for the prediction of polypeptide aggregation propensity. Curr Opin Chem Biol 2006;10:437–44.
[50] Saiki M, Konakahara T, Morii H. Interaction-based evaluation of the propensity for amyloid formation with cross-beta structure. Biochem Biophys Res Commun 2006;343:1262–71.
[51] Galzitskaya OV, Garbuzynskiy SO, Lobanov MY. Is it possible to predict amyloidogenic regions from sequence alone? J Bioinform Comput Biol 2006;4:373–88.
[52] Zhuqing Z, Hao C, Luhua L. Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics 2007;23:2218–25.
[53] Xin L, Ya-Pu Z. Switch region for pathogenic structural change in conformational disease and its prediction. PLoS One, in press.

Please cite this article in press as: Liu X, Zhao Y-P. Simulated pathogenic conformational switch regions matched well with the biochemical findings. J Biomed Inform (2010), doi:10.1016/j.jbi.2009.12.003