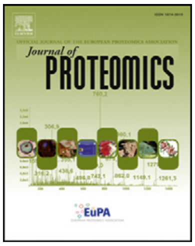
Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot



Review

Innovations in proteomic profiling of cancers: Alternative splice variants as a new class of cancer biomarker candidates and bridging of proteomics with structural biology☆,☆☆

Gilbert S. Omenn[a, b, c, d, e, *], Rajasree Menon[a], Yang Zhang[a, f]

[a]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109-2218, USA
[b]Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109-2218, USA
[c]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109-2218, USA
[d]School of Public Health, University of Michigan, Ann Arbor, MI 48109-2218, USA
[e]Institute for Systems Biology, Seattle, WA 98101, USA
[f]Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109-2218, USA

ARTICLE INFO

Available online 17 April 2013

Keywords:
Splicing
Differential expression
Cancer biomarkers
Computational modeling

ABSTRACT

Alternative splicing allows a single gene to generate multiple RNA transcripts which can be translated into functionally diverse protein isoforms. Current knowledge of splicing is derived mainly from RNA transcripts, with very little known about the expression level, 3D structures, and functional differences of the proteins. Splicing is a remarkable phenomenon of molecular and biological evolution. Studies which simply report up-regulation or down-regulation of protein or mRNA expression are confounded by the effects of mixtures of these isoforms. Besides understanding the net biological effects of the mixtures, we may be able to develop biomarker tests based on the observable differential expression of particular splice variants or combinations of splice variants in specific disease states. Here we review our work on differential expression of splice variant proteins in cancers and the feasibility of integrating proteomic analysis with structure-based conformational predictions of the differences between such isoforms.

This article is part of a Special Issue entitled: From Genome to Proteome: Open Innovations.

Contents

☆ Presented as a Keynote Lecture at the 9th Siena Conference on Proteomics, Siena, Italy, 27 August, 2012.

☆☆ This article is part of a Special Issue entitled: From Genome to Proteome: Open Innovations.

* *Corresponding author at: Computational Medicine & Bioinformatics, Internal Medicine, Human Genetics, and Public Health; University of Michigan, Ann Arbor, MI 48109-2218, USA.* Tel.: +1 734 763 7583; fax: +1 734 615 6553.

E-mail address: gomenn@umich.edu (G.S. Omenn).*URL:* http://www.ccmb.med.umich.edu/omenn (G.S. Omenn).

1874-3919/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jprot.2013.04.007

1. Introduction

In February 2001, *Nature* and *Science* simultaneously published now-classic issues devoted to the sequencing results and early biological applications of the landmark Human Genome Project accomplished by the public and private-sector research teams [1,2]. Five days later, the 21 February issue of *The Financial Times* presented the article shown in Fig. 1 “Searching for the Real Stuff of Life” [3]. Note that the double-helix has been moved into the shadows, off-stage, while the robust globular protein has taken center stage! The article referred to the enormous task to “decipher the human protein set”, the proteome, as “Biotech’s Next Holy Grail”.

The Siena Conferences have been in the forefront of the development of the field of proteomics, even the naming of the field with the term suggested by Marc Wilkins of Australia in 1995. Our theme for this 9th Conference is “From Genome to Proteome”. The overall drivers are these:

- Proteins are the major action molecules of cells
- Proteins and their isoforms are dynamic
- Proteins play critical roles in gene regulation
- Modern instruments, reagents, and bioinformatics facilitate integration and modeling of data from multiple ‘omics platforms
- Proteins are the primary targets of drugs and can be drugs themselves, as well as biomarkers for diagnosis, prognosis, and response to therapy

During the past few months there have been several major science policy reports in the United States that strongly highlighted proteomics:

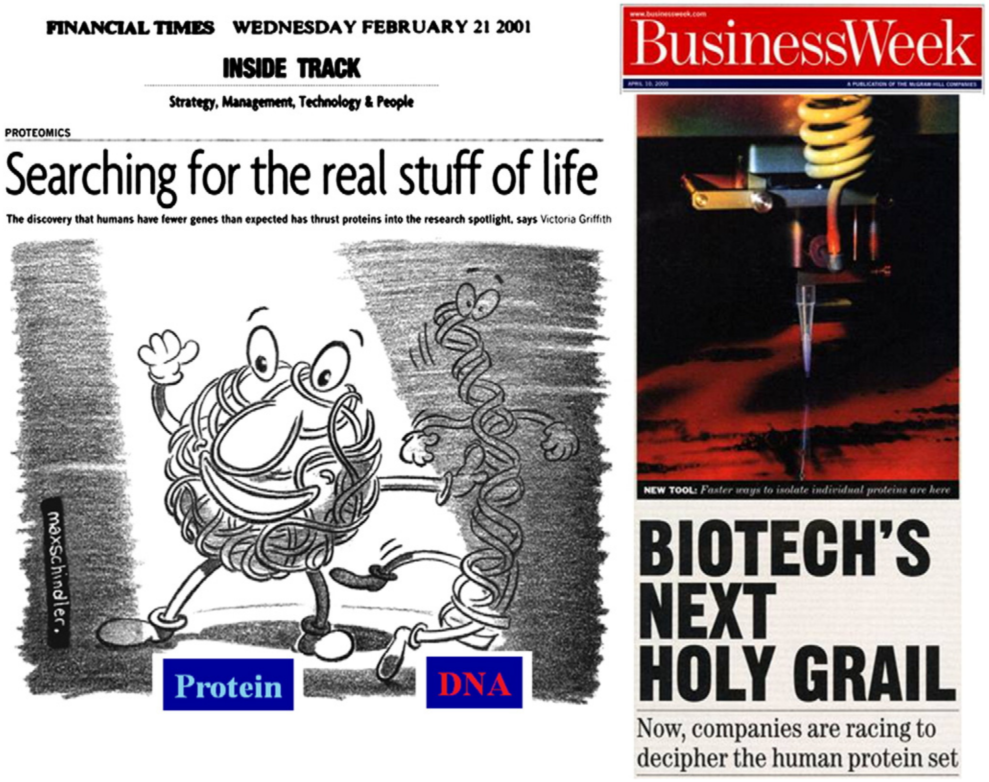
- Vidal, Chan, Gerstein, Mann, Omenn, Tagle, Sechi. The human proteome. Clinical Proteomics 2012 [4]. This report from the NIH Workshop on Human Proteomics emphasized the interactome and the path from biomarker candidate to diagnostic test.
- Hood, Omenn, Moritz, Aebersold, Yamamoto, Amos, Hunter-Cevera, Locascio. Proteomics technologies, a grand challenge in life sciences. Proteomics 2012 [5]. This report from the Gaithersburg Workshop hosted by the National Institute for Standards and Technology addressed the essential role of proteomics in realizing the goals of the Human Genome Project, identified performance challenges and emerging proteomics technologies, and showed applications for health, agriculture and nutrition, energy and environment, and national security.
- Office of Science and Technology Policy. The National Bioeconomy Blueprint, April 2012 [6]. Three “foundational fields” for the coming decade were highlighted: synthetic biology, proteomics, and computational biology.
- Institute of Medicine. Evolution of Translational Omics: Lessons Learned and Path Forward. Micheel, Nass, Omenn (eds). National Academy Press, March 2012 [7]. This report presented a framework for discovery, validation, and clinical utility phases of development of multi-analyte diagnostic tests. Strong recommendations were made for the responsibilities of investigators, lab directors, research institutions, funders, regulators, and journals.

The use of proteomics in cancer biomarker research has two complementary starting points. The first is to directly profile tumor specimens for diagnosis and stratification of patients, for prognosis with or without particular therapies, and for clues to mechanisms and to circulating biomarkers. The second is to profile proteins in the blood plasma to discover and validate biomarkers for earlier or more specific diagnoses and to apply such biomarkers to predict response to treatment and monitor patients for recurrence or metastasis of the tumor.

2. Strategies for biomarker discovery from combined analyses of tumor tissues and plasma

There are now four strategies with high promise for developing tumor-specific and organ-specific biomarkers that can be assayed in the circulation:

1. Start with microarray or next-gen sequencing evidence for carcinogenic pathway mechanisms in tumor and track corresponding protein biomarker candidates to the plasma. This is a major strategy at the Institute for Systems Biology, identifying differentially-expressed transcripts and proteins



FINANCIAL TIMES WEDNESDAY FEBRUARY 21 2001
INSIDE TRACK
Strategy, Management, Technology & People
PROTEOMICS

Searching for the real stuff of life

The discovery that humans have fewer genes than expected has thrust proteins into the research spotlight, says Victoria Griffith

maxSchindler.
Protein
DNA
www.businessweek.com
BusinessWeek
APRIL 10, 2000
A PUBLICATION OF THE McGRAW-HILL COMPANIES
NEW TOOL: Faster ways to isolate individual proteins are here

BIOTECH'S NEXT HOLY GRAIL

Now, companies are racing to decipher the human protein set

Fig. 1 – "Searching for the Real Stuff of Life". Cartoon and story from the Financial Times of London, p.14, 21 February 2001, showing the central role of the proteome.

in brain or in liver, then searching for the corresponding proteins in the plasma [8]. In the case of prostate cancers, the common TMPRSS2/ETS fusion protein and the metabolite associated with metastasis, sarcosine, were discovered in tumor tissue and cell lines and then converted into assays in urine by the Chinnaiyan lab at the University of Michigan [9,10].

2. Perform targeted proteomics with SRM/MRM to identify and quantify these candidates. The transition from shotgun analyses or transcript expression information to targeted analysis of candidate proteins is a major development. For example, Huttenhain et al. performed SRM of 1172 candidate proteins and detected 182 in immunodepleted plasma and 408 in urine specimens from ovarian cancer patients [11].
3. Detect auto-antibodies in plasma as a biological amplification of tumor protein signals, then confirm in tumor tissue. There is no laboratory method for proteins like the polymerase-chain reaction (PCR) which has transformed the sensitivity and specificity of analysis of nucleic acids. Nature, however, does provide an amplification in the form of the immune response, which generates concentrations of antibodies two, three, or four orders of magnitude greater than the circulating concentrations of the corresponding tumor antigen.
4. Identify alternative splice isoforms of biologically meaningful proteins in cancers and in plasma of humans and mouse models. This new approach is the heart of this Review article.

3. Alternative splice variants of proteins represent both a source of molecular diversity and a new class of biomarker candidates

One of the most remarkable developments in biological evolution is the emergence of gene structures with exons and introns and a complex splicing machinery in cells that processes heterogeneous nuclear RNAs and generates several different mRNA and protein products from individual genes. Just describing gene or protein expression as "up-regulated" or "down-regulated" ignores the fact that these transcripts and proteins are mixtures. As shown below, these splice variants can and often do have dramatically different functions; when the proteins fold and compete similarly for target sites, they may, in fact, have opposing actions, such as pro-apoptotic and anti-apoptotic activities.

Alternative splicing generates protein diversity without increasing genome size. This phenomenon seems to explain how humans can "get by" with only 20,000 protein-coding genes, whereas there were predictions of 50,000 to 100,000 or more protein-coding genes when the Human Genome Project was launched. The splice variants cannot be identified in genome sequences, but the splicing can be mapped to the gene exon/intron structures. Thus, as shown in Fig. 2, we can deduce from the peptide and gene sequences the kinds of splicing events, including alternative 5′ or 3′ start sites, mutually exclusive exons (exon swaps), intron retention, alternative promoters, and alternative polyadenylation. There are examples of every kind

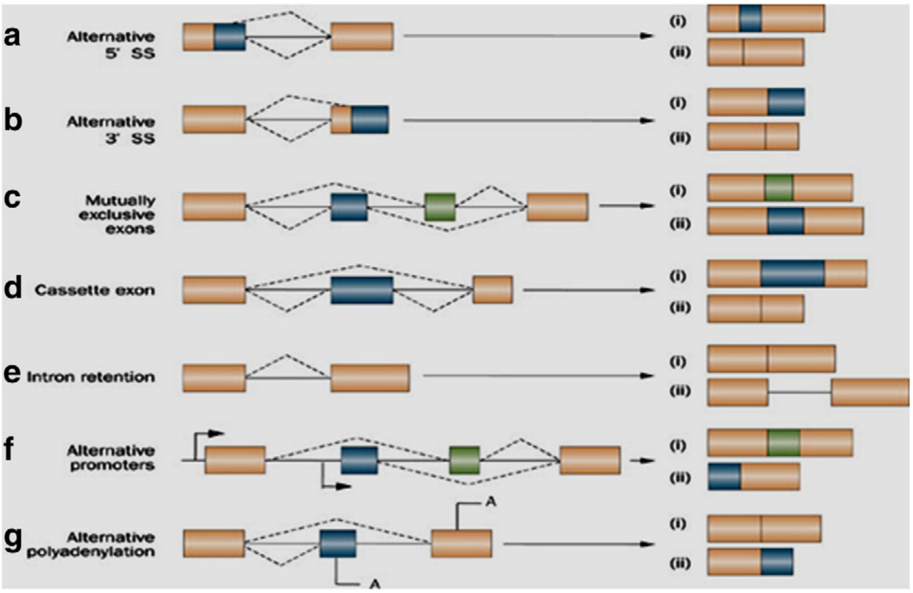


Fig. 2 – Seven different mechanisms of alternative splicing. From Rajan, P., et al, Nature Reviews Urology, 2009, 6: 454–460, with permission (requested).

of splicing in cancers. Splice events that affect the protein-coding region of the mRNA generate proteins differing in sequence and functions; splicing within the non-coding regions can alter regulatory elements such as translation enhancers or RNA stability domains, which in turn influence protein expression. There are several databases which present alternatively spliced transcripts [12,13].

3.1. *Differential expression and altered functions of splice variants*

In recent years, reports on distinct expression patterns and functions of the splice variants of a gene in normal or disease conditions have become frequent. Smith et al. [14] showed differential expression of two kcnq2 splice variants. Kcnq2 is a brain-derived gene involved in neuronal M current. The long variant is preferentially expressed in differentiated neurons, whereas the short transcript is prominent in fetal brain, undifferentiated neuroblastoma cells, and brain tumors. The long kcnq2 variant, transfected into mammalian cells, produces a more slowly activating voltage-gated K(+) current; co-transfection of the shorter variant with the longer variant results in attenuation of the K(+) current [14]. Liu et al. [15] reported the specific functions of the splice variants of dynamin2 (dyn2). Dynamins are multifunctional GTPases involved in endocytosis, intracellular trafficking, cell signaling, and cytokinesis. Although all four dyn2 splice variants could equally restore clathrin-mediated endocytosis, dyn2ba and dyn2bb were more effective at restoring p75 exocytosis [15]. Alternatively-spliced transcripts of the osr2 gene encode osr2-L (312 aa) and osr2-S (276 aa), which have opposite transcriptional activities, activation and repression [16]. The difference between the variants is in the C-terminal region translated from the third exon of the transcripts (60 aa in osr2-L and 24 aa in osr2-S). Osr2-L contains five C2H2 zinc finger domains, including two domains in the C-terminal spliced region, compared to three total C2H2 zinc finger domains in osr2-S. The opposing functions of these two variants seem to result from different numbers of zinc-finger domains, with different phosphorylation patterns and/or different affinity for DNA-binding [16]. The ratio of the expression of splice variants can play a significant role in the normal functioning of a biological system. For example, in the normal human brain, the ratio of tau-4R to tau-3R is approximately 1; this balanced isoform ratio appears to be essential for proper neuronal function [17,18]. The splicing event, which results in exon 10 inclusion or skipping, gives rise to tau-4R or tau-3R, respectively. Finally, Sevcik et al reported an alternative splice variant delta14-15 of BRCA1 with an in-frame deletion of part of the regulatory serine-containing domain, which impairs DNA double-strand break repair capacity in MCF-7 breast cancer cells in vitro [19].

3.2. *Role of splice variants in human cancers*

Studies have emerged showing the involvement of specific gene splice variants in different types of cancers or cancer-related processes. For example, the Nek2C splice variant of the serine/threonine kinase Nek2 is involved in breast cancer development; Nek2C inhibition may be a potential therapeutic approach to targeting some types of human breast tumors [20]. Specific ligand binding to the receptor for advanced glycation end-products (RAGE) can activate signal transduction pathways which may be involved in many degenerative diseases and cancers [21]. Lertwittayapon et al. showed that RAGE variant 1 (RAGEv1), a major soluble form of RAGE in the circulating blood, can neutralize deleterious ligands, thus diminishing signaling that might lead to inflammation in cancers. They proposed that this variant could provide a potential alternative therapy for the treatment of liver cancer [21]. Another interesting example is the role of carboxypeptidase E (CPE) in Wnt signal transduction pathways, especially implicated in colorectal cancer [22]. Skalka et al showed that a splice variant form of CPE activates the Wnt signaling pathway, whereas the full length canonical CPE is an inhibitor of Wnt/B-catenin signaling [22].

3.3. The University of Michigan modified ECgene database of potential translation products

We began our search for splice variants from mass spectrometry datasets of mouse and human proteomes, and confirmed peptide/protein variants using qRT-PCR analysis of the mRNA. Now we can begin with much more comprehensive RNA-Seq datasets and, when feasible, determine whether those splice transcripts are translated into splice variant proteins.

We combined Ensembl with ECgene data translated in three frames to create a modified ECgene database with all potential protein sequences, totaling 10.4 million entries for the mouse and 14.2 million entries for the human. See Menon et al and Omenn & Menon for details [23,24]. mzXML files containing the experimental mass spectra information are searched against this database using X!Tandem software. A collection of common contaminants was added and a set of reversed sequences was generated for each analysis to estimate false discovery rates. Peptides were integrated into a list of proteins using TransProteomic Pipeline and/or the Michigan Peptide-to-Protein Integration workflow. Peptides with <1% false discovery rate were used for the analysis. To characterize the splice variant proteins, we used InterProScan, MotifScan, Gene Ontology, and FuncAssociate, and displayed protein–protein interactions with the Cytoscape plug-in for Michigan Molecular Interactions (MiMI), a consolidated resource of six different databases for protein-protein interactions [25].

3.4. Identification of splice variant peptides in the plasma of mice with pancreatic ductal adenocarcinoma (DePinho/Bardeesy model)

The Kras^{G12D} activation/Ink4a/Arf deletion model of pancreatic ductal adenocarcinoma was genetically engineered by DePinho and Bardeesy to match the molecular lesions of human PDAC; it recapitulates the histologic progression and clinical features of the human disease [26]. We exploited this model to test the hypothesis that cancer-specific splice variants could be identified in MS analyses of plasma proteins from mice with tumors carrying these molecular lesions, compared with wild-type mice [24]. After immunodepletion of the three most abundant proteins — albumin, immunoglobulins, and transferrin, and D3 vs. D0-acrylamide labeling of cysteine residues, the combined tumor and wild-type plasma samples were fractionated into a total of 163 fractions, digested with trypsin, and analyzed with a ThermoFinnigan LTQ-FT mass spectrometer.

Our integrated analysis revealed 420 distinct splice isoforms; 92 were novel, not matching any previously annotated mouse protein sequences. For seven of these novel variants, we prepared primers and validated the predicted sequences in the mRNA using qRT-PCR. The acrylamide labeling permitted relative quantitation of 28 of the 92 novel proteins (those containing cysteines). We visualized the splicing events using the UCSC Genome Browser. We demonstrated differential expression for peptides from muscle-type pyruvate kinase, malate dehydrogenase 1, glyceraldehyde-3-phosphate dehydrogenase, proteoglycan 45, minichromosome maintenance complex component 9, high mobility group box 2, and hepatocyte growth factor activator. We presented literature evidence that many of these splice variants are probably involved in pancreatic cancers, including alpha-fetoprotein, apolipoprotein E, ceruloplasmin, fibronectin, glyceraldehyde-3-phosphate dehydrogenase, hemopexin, peptidyl-prolylisomerase, and tubulin alpha among the novel splice variants, and acyl coA acetyl-transferase, chromograinin b, granulin, insulin-like growth factor binding protein 2, and regenerating islet-derived 3alpha among the known variants that also had significant up-regulation in plasma of the tumor-bearing mice.

One of the most interesting proteins is pyruvate kinase, the critical enzyme in the metabolic switch to aerobic glycolysis in cancers known since 1929 as “the Warburg effect” [27]. We discuss the structural consequences of the exon swap of exons 9 and 10 later in this paper.

Another interesting differentially-expressed splice variant is high mobility group box 2, which is involved in DNA repair; its gene is located at 4q32-34, a region associated with familial pancreatic cancer [28]. We also searched our peptide findings for variants of proteins chosen as potential pancreatic cancer biomarkers in a parallel study of this same mouse model [29]; we found variants of three of the nine candidate proteins assayed by ELISA in humans: lipocalin 2 (LCN2), regenerating islet-derived 3 (REG3A), and tumor necrosis factor receptor superfamily member 1A (TNFRSF1A). These three proteins were included in a panel of five proteins that discriminated between stored serum specimens from 13 participants in the CARET lung cancer chemoprevention trial [30] who 7–13 months later were diagnosed with pancreatic cancer and 13 matched controls who did not develop cancer in a subsequent four-year follow-up period [29].

3.5. Identification of splice variant peptides in tumor tissue of mice with Her2/neu-amplified breast cancer (chodosh model)

This study analyzed tumor and normal mammary tissue LC-MS/MS datasets from the Chodosh mouse model of Her2/neu-driven breast cancer, which accounts for 15–20% of breast cancers in humans [31].

We found a total of 608 distinct alternative splice variants, 540 known and 68 novel [32]. There were 216 more from the tumor lysate than from the normal sample (505 vs 289), probably reflecting greater cellularity and higher expression per cell. We chose 32 of the 45 novel proteins expressed only in tumor specimens for confirmation with qRT-PCR; all were confirmed except for one primer which did not work, and 29 of 31 showed increased mRNA expression. Of the 15 biomarker candidates Whiteaker et al. [31] confirmed as over-expressed in tumor lysates with MRM-MS, we found that 10 had splice variants in our analysis; of course, we had no information on the functional activities of the different isoforms of these or any other proteins from proteomics analyses.

Among the 68 novel proteins we demonstrated variants resulting from new translation start sites, new splice sites, extension or shortening of exons, deletion or switch of exons, retention of introns, and translation in an alternative reading frame. Our annotations revealed multiple variants with potential

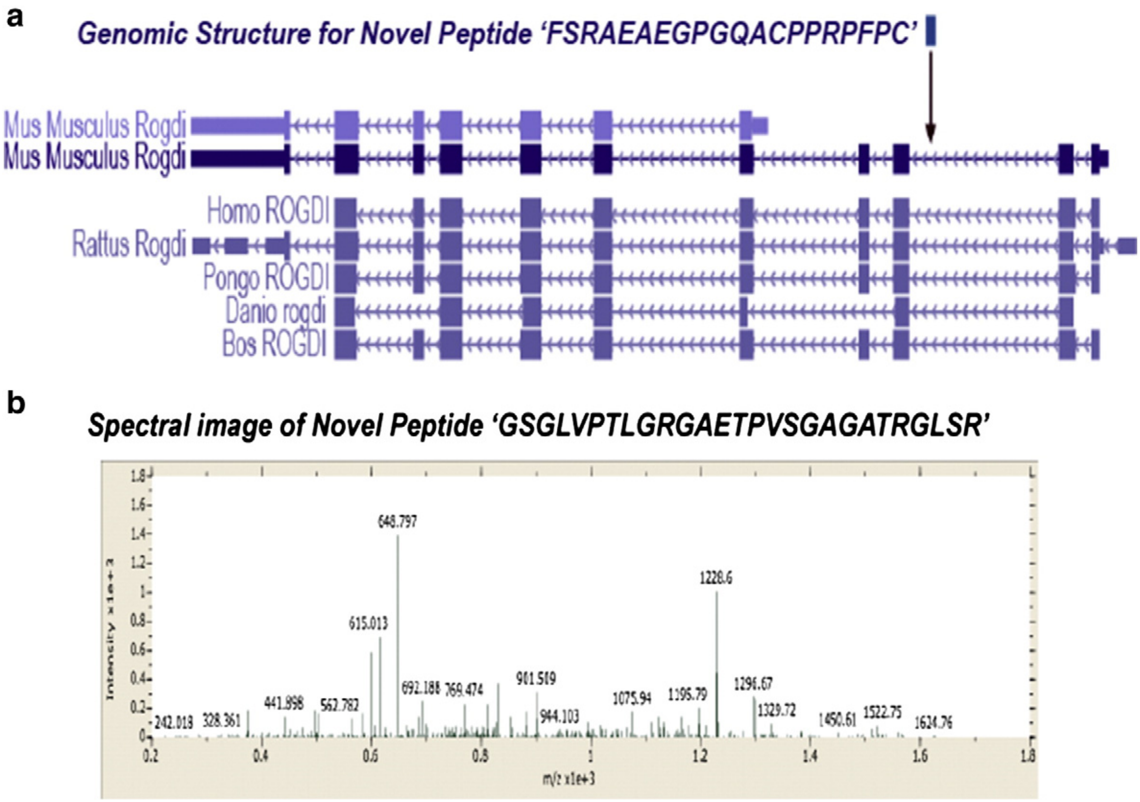


Fig. 3 – Two splice variant proteins, Rogdi and Sox 7, which interact with the carboxy-terminal region (BRCT) of the BRCA1 breast cancer gene. (a) Genomic structure of the Rogdi gene as shown on the University of California Santa Cruz (UCSC) Genome Browser. The novel peptide aligns to an intronic region. (b) Spectral image using Insilicos Viewer for the novel peptide from the transcription factor Sox 7.

significant functional motifs, including two relating to BRCA1 through binding to its BRCT domain.

The peptide sequence 'FSRAEAEGPGQACPPRPFPC' is in the second intronic region of leucine-zipper-containing LF (Rogdi) gene (Fig. 3a). Using Splice Site Prediction by Neural Network from the Berkeley Drosophila Genome Project (http://w.fruitfly.org/seq_tools/splice.html), we found a predicted donor splice site 'gactgaggtgaggtg' where the novel peptide was identified as the coding sequence with a splice site prediction score of 0.93. Functional motifs identified in this section of intronic sequence include LIG_BRCT_BRCA1_1, a phosphopeptide motif which interacts directly with the carboxy-terminal domain of BRCA1. The second case involves the peptide 'GSGLVPTLGRGAETPVSG AGATRGLSR', aligned to the first intronic region of transcription factor *sox7*; the very same LIG-BRCT_BRCA1_1 motif was found in this intronic region (Fig. 3b). A colleague working on BRCA-related mechanisms is exploring experimentally these predictions. In addition, the structures of these protein-protein interactions can be modeled by computer-based methods including protein docking and recently developed multiple-chain threading and reassembly algorithms [33].

3.6. Ongoing splice variant studies with human cancer cell lines

With the advent of high-throughput RNA-sequencing, we have begun examining both RNA and proteins for evidence of biologically interesting splice variants. Current studies include combined analysis of the Her2/neu+ (ERBB2+) breast cancer cell line SKBR3, and six other ERBB2+ gastric breast, or colon cancer cell lines, as part of the Chromosome 17 project of the Human Proteome Project [34]. We also are investigating the VCaP prostate cancer cell line and the normal prostate cell line RWPE and the SUM149 inflammatory breast cancer cell line.

4. Bridging protein chemistry/structural biology and proteomics with computational modeling of proteins

Experimentally determined structures of protein splice isoforms are rare; in fact, there are only 7 full-length pairs of such isoforms in the enormous Protein Data Bank (PDB) and the Alternative Splicing and Transcript Diversity (ASTD) database. Homology modeling methods are poor at predicting atomic-level structural differences because of the high sequence identity between the isoforms. We have exploited the state-of-the-art protein structure prediction method I-TASSER [35] to analyze the folding, conformation, and likely functional consequences of alternative splicing of proteins identified in the Her2/neu-induced breast cancer model described above [36].

The I-TASSER algorithm was designed to construct full-length protein models by reassembling the continuous structural fragments excised from the protein templates as identified by the multiple threading technique [35]. The Monte Carlo structural assembly simulations in I-TASSER are driven by optimized knowledge and physics-based force-field analyses. A major advantage of I-TASSER over traditional homology modeling is its ability to refine the template structure closer to the native state, which is critical to success in modeling spliced isoforms. The I-TASSER-based methods have been ranked at the top in the past four biennial international competitions for Critical Assessment of Techniques for Prediction of Structures of Proteins (CASP) [37].

Based on the I-TASSER modeling, we have demonstrated that its attributes in *ab initio* structural assembly and template refinement can partially differentiate atomic details of splice protein variant pairs [36]. First, we benchmarked the approach with all seven pairs of protein splice isoforms with solved structures in PDB, which resulted in structural models with an average RMSD = 1.72 Å to the native after excluding all homologous templates to the targets. Most of the structural variations in the isoform pairs were due to exon swapping. Even alternative splice variants whose structures are very similar may have functional differences due to absence of a functionally critical residue or altered post-translational modifications of residues in the swapped exon. For example, in the case of acid phosphatase (acp1) variants, the Mg^{2+} binding site is missing in the 1xwwA variant.

In the second step, we used the strategy to model three cancer-related variant pairs reported to have opposite functions, but lacking experimentally-derived structures: Bcl-x, caspase 3, and odd-skipped related 2. In each isoform pair, we observed structural differences in regions where the presence or absence of a motif can directly influence the distinctive functions of the variants. For example, an additional 63 amino acids (aa 129–191) create an extra domain in the core structure of bclx-L (233 aa) compared with bclx-S (170 aa); the shorter variant is missing the two Bcl-2 family motifs BH1 and BH2 while the longer variant contains all four Bcl-2 homology motifs (BH1-4). This difference results in completely different topology and function; bclx-L is anti-apoptotic, while bclx-S is pro-apoptotic.

Then we applied I-TASSER to five splice variant pairs over-expressed in the mouse Her2/neu mammary tumor we had studied: annexin 6, calumenin, cell division cycle 42 (cdc42), polypyrimidine tract binding protein 1 (ptbp1), and tax1-binding protein 3 (tax1bp3). These pairs were chosen based on the following five criteria: differential expression, annotated as a known protein in Ensembl, at least 75% sequence identity with the canonical protein, known homologous variants of the protein pair in Homo sapiens, and an I-TASSER confidence score (C-score) for both the variants >– 1.5 to ensure the quality of structure prediction. Despite the high sequence identity between the variant pairs (99, 92, 95, 95, 79%, respectively), structural differences were revealed in biologically important regions of these protein pairs.

For example, the only difference between anxa6-001 and anxa6-002 at the sequence level is the presence of six residues in anxa6-001 (VAAEIL, aa 525–530) that are missing in anxa6-002. The global topology of the I-TASSER models of the two isoforms is almost identical, with RMSD = 0.38 A and TM-score = 0.99. However, there is an obvious local structural change in the region due to the absence of "VAAEIL" residues (aa 525–530 in anxa6-001), as identified by TM-align [38]. As reported, these six residues are in the end of a helical region (blue-colored in the original figure) which is followed by a loop. Because of the absence of the six residues, the loop is smaller in the shorter variant. The nearby proline-directed kinase phosphorylation ([ST]P) site followed by a serine phosphorylation site moves from 535–537 to 529–531, inside the helix region in anxa6-002, where phosphorylation is less probable than for anxa6-001.

The new I-TASSER models in Fig. 4 show that the threonine and proline residues are buried by other atoms in the anxa6-002 variant whereas the hydroxyl group of serine (S531 in anxa6-002)

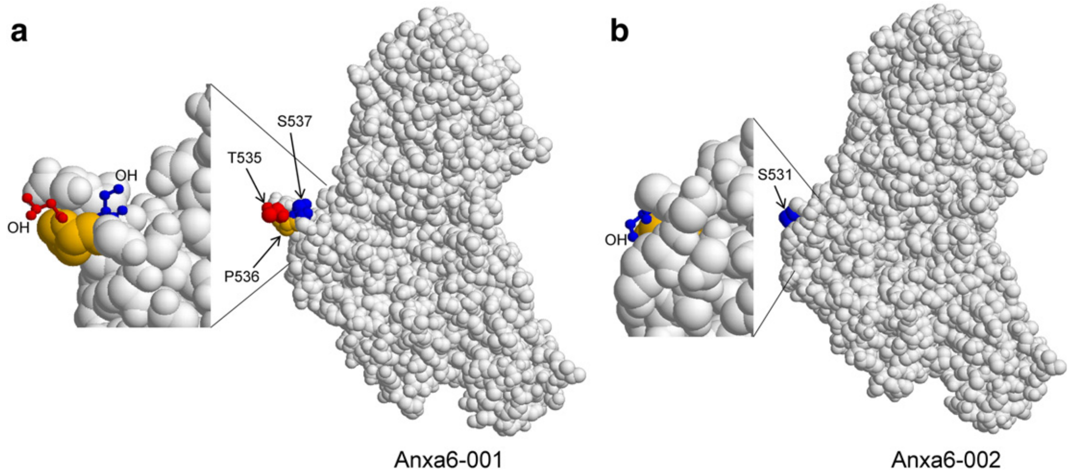


Fig. 4 – The I-TASSER models for two splice isoforms of Annexin 6. (a) The "TPS" residues in anxa6-001 are exposed to solvent, which helps increase the likelihood of phosphorylation in the post-translational modification. The hydroxyl groups, which are the target of kinases for phosphorylation, are highlighted in the inset. (b) Due to the absence of "VAAEIL" residues (aa 525–530 in anxa6-001) in the anxa6-002 variant, the 'TPS" residues are either partially or completely buried by other atoms which significantly reduces the possibility of phosphorylation of these residues; note, however, that serine-531 remains accessible for phosphorylation. As described in the text, this phosphopeptide was identified experimentally.

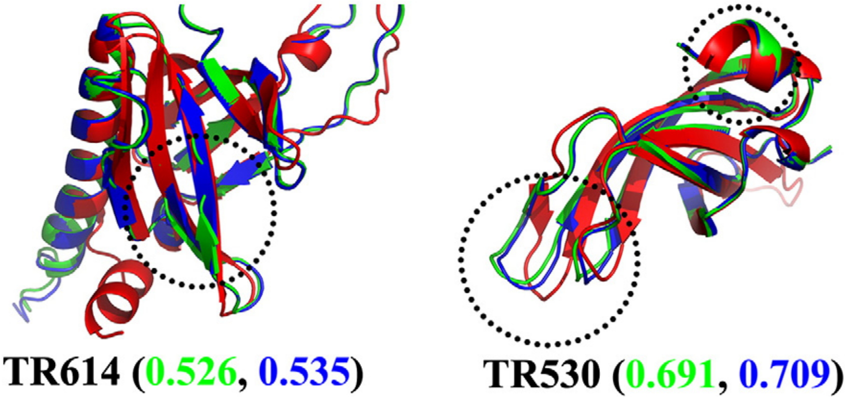


Fig. 5 – Two successful examples of atomic-level structural refinement by fragment-guided molecular dynamic simulations for target proteins TR614 and TR530. The global distance test score (GDT-HA) increases by 9 and 8 units, respectively. Red, green and blue represent native structure, initial model, and refined model with dotted circles highlighting the regions of more pronounced structural refinements. Values in parentheses indicate GDT-HA score compared to the native (same color code).

remains accessible for phosphorylation (see inset of Fig. 4B). In order to search for phosphopeptides from the spliced region of anxa6, we performed a fresh analysis of the mass spectrometric data with our custom database using X! Tandem, specifying phosphorylation on serine or threonine (phospho(S) and phospho(T)) as potential residue modifications. Since phosphorylation is usually present at low stoichiometry and our dataset was not enriched for phosphopeptides, it was striking that we identified a spectrum from the normal sample that matched to the peptide 'DQAQEDAQVAAEILEIADTPSGDKTSLETR' with 3281.506 daltons as the calculated peptide mass plus a proton (mh). The unmodified mh of this peptide is 3201.539 daltons; the additional 79.967 daltons can be accounted for precisely by phosphorylation of either the threonine or serine residue in the peptide. We did not find such a phosphopeptide from the tumor sample. However, we did find multiple high quality spectra from the tumor sample that identified the sequence 'DQAQEDAQEI ADTPSGDKTSLETR', the unique peptide that matches the anxa6 short variant (with residues 'VAAEIL' missing). None of these spectra revealed modification by phosphorylation. Even though the phosphopeptide from the unique region of the long anxa6 variant was identified with only a single spectrum in the X! Tandem search, these observations are consistent with our functional inference from the structural comparison of the anxa6 variants [35] that the longer anxa-001 variant is more prone to undergo phosphorylation at Thr-535 or Ser-537 than is the anxa-002 variant at the Thr-529 or Ser 531 sites. Post-translational phosphorylation of anxa6 has been reported to be associated with cell growth in 3T3 fibroblasts and human T-lymphoblasts [39]; we previously predicted that the critical phosphorylation may occur at Thr-535 and/or Ser-537 in the loop region. We have now strikingly refined this prediction, which we hope experimentalists will test.

Interesting features of the other four splice variant pairs were revealed, as well, by the I-TASSER analysis [36]. This novel approach of functional inferences for alternative splice variants by comparing their predicted structures has provided insights to explore the roles of these isoforms in the complex mechanisms of various cancers or other diseases.

5. New *ab initio* protein folding and refinement algorithms are essential to understanding structural and functional consequences of RNA alternative splicing

The key to the success of I-TASSER modeling of the alternatively spliced protein isoforms is the ability to model the global and local structural changes induced from the local residue variations along the sequences. Nevertheless, since I-TASSER simulations start from the threading alignments, it is often challenging for I-TASSER to accurately differentiate the subtle atomic-level structural variations if threading programs identify highly similar templates along the regions. To alleviate these difficulties, we recently developed a new approach, QUARK [40], for *ab initio* protein folding which assembles protein folds from small fragments and without using global template structures. QUARK was ranked as the best *ab initio* folding algorithm in CASP9 for template-free modeling [41]. In general, a combination of the template-based and *ab initio* folding approaches, e.g. using template-based modeling for global structure construction and *ab initio* folding to assemble the specific alternatively spliced regions, should represent a promising strategy to high-resolution structural prediction of protein splice isoforms. The development of such a hybrid approach is underway [42,43].

For accelerating the folding simulations, both I-TASSER and QUARK are based on reduced modeling, i.e. residues are represented by the side-chain center and Cα (in I-TASSER) or backbone heavy atoms (in QUARK). Thus, full atomic refinements are usually needed to repack the loops and side-chain rotamers following the global fold constructions. In the recent development of fragment-guided molecular dynamics simulations (FG-MD) [44,45], we proposed to use the low-resolution model as a probe to identify fragment analogs from the PDB. The distance maps, together with backbone-oriented hydrogen bonding, are then used to guide the simulated annealing MD simulations. The protocol was tested on 181 proteins of 66 to 222 residues. It was found that structure models with correct folds with TM-score >0.5 can often be pulled closer to native, but improvement for the models of incorrect folds (TM-score < 0.5) is much less pronounced. These data indicate that fragment distance maps essentially re-shaped the MD energy landscape and improved the funnel shape in the targets with a radius of TM-score ~0.5 [44]. In Fig. 5, we show two typical examples of FG-MD target refinements in CASP9 (TR614 and TR530), where FG-MD is the only group in the Refinement Section which could drive models closer to the native by having an average GDT-HA (global distance test/high accuracy) score higher than the initial models [46].

The major goal of the isoform structure determinations is to understand the functional implications of alternative splicing in living cells. To predict the biological functions of protein molecules, we developed COFACTOR [47,48], based on the sequence-to-structure-to-function paradigm [35]. Under the assumption that proteins of similar structures have similar functions, the predicted I-TASSER structural models are matched with known proteins in the representative function library, BioLiP [49], based on both global and local structure comparisons. The function can be inferred from template proteins if they have similar global topology or local pocket geometry around the active/binding sites. The COFACTOR algorithm was tested in the 2010 CASP9 experiment and generated the most accurate binding site predictions based on both Z-score and Matthews correlation coefficient (MCC) [50].

Overall, such efforts integrating computational protein structure and function modeling represent an emerging important avenue toward understanding protein-level implications of RNA alternative splicing in living cells.

Acknowledgment

This work was supported by grants. GSO acknowledges support from NIH grants RM-08-029; P30U54ES017885; U54DA021519; and UL1RR24986. YZ is supported in part by the NSF Career Award (DBI 0746198), and the National Institute of General Medical Sciences (GM083107, GM084222).

REFERENCES

[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

[2] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science 2001;291: 1304–51.

[3] Proteomics: Searching for the real stuff of life.The Financial Times; 2001. p. 14 [21 Feb 2001].

[4] Vidal M, Chan DW, Gerstein M, Mann M, Omenn GS, Tagle D, et al. The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. Clin Proteomics 2012;9:6.

[5] Hood LE, Omenn GS, Moritz RL, Aebersold R, Yamamoto KR, Amos M, et al. New and improved proteomics technologies for understanding complex biological systems: addressing a grand challenge in the life sciences. Proteomics 2012;12: 2773–83.

[6] U.S. Office of Science and Technology Policy. National Bioeconomy Blueprint. Washington: The White House; 2012. p. 43.

[7] Micheel C, Nass SJ, Omenn GS. Evolution of translational omics lessons learned and the path forward. Washington, D.C.: National Academies Press; 2012. p. 338.

[8] Qin S, Zhou Y, Lok A, Tsodikov A, Yan X, Gray L, et al. SRM Targeted proteomics in search for biomarkers of HCV-induced progression of fibrosis to cirrhosis in HALT-C patients. Proteomics 2012;12:1244–52.

[9] Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8.

[10] Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 2009;457: 910–4.

[11] Huttenhain R, Soste M, Selevsek N, Rost H, Sethi A, Carapito C, et al. Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. Sci Transl Med 2012;4:142ra94.

[12] Larsson TP, Murray CG, Hill T, Fredriksson R, Schioth HB. Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. FEBS Lett 2005;579:690–8.

[13] Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S. ECgene: genome annotation for alternative splicing. Nucleic Acids Res 2005;33: D75–9.

[14] Smith JS, Iannotti CA, Dargis P, Christian EP, Aiyar J. Differential expression of kcnq2 splice variants: implications to m current function during neuronal development. J Neurosci 2001;21:1096–103.

[15] Liu YW, Surka MC, Schroeter T, Lukiyanchuk V, Schmid SL. Isoform and splice-variant specific functions of dynamin-2 revealed by analysis of conditional knock-out cells. Mol Biol Cell 2008;19:5347–59.

[16] Kawai S, Kato T, Inaba H, Okahashi N, Amano A. Odd-skipped related 2 splicing variants show opposite transcriptional activity. Biochem Biophys Res Commun 2005;328:306–11.

[17] Goedert M, Jakes R. Expression of separate isoforms of human tau protein: correlation with the tau pattern in brain and effects on tubulin polymerization. EMBO J 1990;9:4225–30.

[18] Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, et al. Association of missense and 5′-splice-site mutations in tau with the inherited dementia FTDP-17. Nature 1998;393:702–5.

[19] Sevcik J, Falk M, Kleiblova P, Lhota F, Stefancikova L, Janatova M, et al. The BRCA1 alternative splicing variant Delta14–15 with an in-frame deletion of part of the regulatory serine-containing domain (SCD) impairs the DNA repair capacity in MCF-7 cells. Cell Signal 2012;24:1023–30.

[20] Liu Z, Wang Y, Wang S, Zhang J, Zhang F, Niu Y. Nek2C functions as a tumor promoter in human breast tumorigenesis. Int J Mol Med 2012;30:775–82.

[21] Lertwittayapon T, Tencomnao T, Santiyanont R. Inhibitory effect of alternatively spliced RAGEv1 on the expression of NF-kB and TNF-alpha in hepatocellular carcinoma cells. Genet Mol Res 2012;11:1712–20.

[22] Skalka N, Caspi M, Caspi E, Loh YP, Rosin-Arbesfeld R. Carboxypeptidase E: a negative regulator of the canonical Wnt signaling pathway. Oncogene 2012;32:2836–47.

[23] Omenn GS, Yocum AK, Menon R. Alternative splice variants, a new class of protein cancer biomarker candidates: findings in pancreatic cancer and breast cancer with systems biology implications. Dis Markers 2010;28:241–51.

[24] Menon R, Zhang Q, Zhang Y, Fermin D, Bardeesy N, DePinho RA, et al. Identification of novel alternative splice isoforms of circulating proteins in a mouse model of human pancreatic cancer. Cancer Res 2009;69:300–9.

[25] Gao J, Ade AS, Tarcea VG, Weymouth TE, Mirel BR, Jagadish HV, et al. Integrating and annotating the interactome using the MiMI plugin for cytoscape. Bioinformatics 2009;25:137–8.

[26] Aguirre AJ, Bardeesy N, Sinha M, Lopez L, Tuveson DA, Horner J, et al. Activated Kras and Ink4a/Arf deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma. Genes Dev 2003;17:3112–26.

[27] Hitosugi T, Kang S, Vander Heiden MG, Chung TW, Elf S, Lythgoe K, et al. Tyrosine phosphorylation inhibits PKM2 to promote the Warburg effect and tumor growth. Sci Signal 2009;2:ra73.

[28] Earl J, Yan L, Vitone LJ, Risk J, Kemp SJ, McFaul C, et al. Evaluation of the 4q32-34 locus in European familial pancreatic cancer. Cancer Epidemiol Biomarkers Prev 2006;15:1948–55.

[29] Faca VM, Song KS, Wang H, Zhang Q, Krasnoselsky AL, Newcomb LF, et al. A mouse to human search for plasma proteome changes associated with pancreatic tumor development. PLoS Med 2008;5:e123.

[30] Omenn GS, Goodman GE, Thornquist MD, Balmes J, Cullen MR, Glass A, et al. Effects of a combination of beta-carotene and vitamin A on lung cancer and cardiovascular disease. N Engl J Med 1996;334:1150–5.

[31] Whiteaker JR, Zhang H, Zhao L, Wang P, Kelly-Spratt KS, Ivey RG, et al. Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. J Proteome Res 2007;6:3962–75.

[32] Menon R, Omenn GS. Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. Cancer Res 2010;70:3440–9.

[33] Mukherjee S, Zhang Y. Protein–protein complex structure predictions by multimeric threading and template recombination. Structure 2011;19:955–66.

[34] Liu S, Im H, Bairoch A, Cristofanilli M, Chen R, Dalton S, et al. A Chromosome-Centric Human Proteome Project (C-HPP) to characterize the sets of proteins encoded in chromosome 17. J Proteome Res 2013;12:45–57.

[35] Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–38.

[36] Menon R, Roy A, Mukherjee S, Belkin S, Zhang Y, Omenn GS. Functional implications of structural predictions for alternative splice proteins expressed in Her2/neu-induced breast cancers. J Proteome Res 2011;10:5503–11.

[37] Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A. Critical assessment of methods of protein structure prediction-Round VIII. Proteins-Struct Funct Bioinform 2009;77:1–4.

[38] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33: 2302–9.

[39] Moss SE, Jacob SM, Davies AA, Crumpton MJ. A growth-dependent post-translational modification of annexin VI. Biochim Biophys Acta 1992;1160:120–6.

[40] Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80:1715–35.

[41] Kinch L, Yong Shi S, Cong Q, Cheng H, Liao Y, Grishin NV. CASP9 assessment of free modeling target predictions. Proteins 2011;79(Suppl. 10):59–73.

[42] Xu D, Zhang J, Roy A, Zhang Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 2011;79(Suppl. 10):147–60.

[43] Zhang Y, Xu D, Yang J, Roy A, Yan R. Protein structure predictions by a combination of I-TASSER and QUARK pipelines. CASP10 Abstract, 2013, p. 248–50.

[44] Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19:1784–95.

[45] Elber R. Progress at last. Structure 2011;19:1725.

[46] MacCallum JL, Perez A, Schnieders MJ, Hua L, Jacobson MP, Dill KA. Assessment of protein structure refinement in CASP9. Proteins 2011;79(Suppl. 10):74–90.

[47] Roy A, Zhang Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. Structure 2012;20:987–97.

[48] Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 2012;40:W471–7.

[49] Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res 2012, http://dx.doi.org/10.1093/nar/gks966.

[50] Schmidt T, Haas J, Cassarino TG, Schwede T. Assessment of ligand-binding residue predictions in CASP9. Proteins 2011;79(Suppl. 10):126–36.