www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 18 May 2016
Accepted: 15 December 2016
Published: 23 January 2017

Novel Bioinformatics–Based Approach for Proteomic Biomarkers Prediction of Calpain-2 & Caspase-3 Protease Fragmentation: Application to βII-Spectrin Protein

Atlal El-Assaad[1], Zaher Dawy[1], Georges Nemer[2] & Firas Kobeissy[2]

The crucial biological role of proteases has been visible with the development of degradomics discipline involved in the determination of the proteases/substrates resulting in breakdown-products (BDPs) that can be utilized as putative biomarkers associated with different biological-clinical significance. In the field of cancer biology, matrix metalloproteinases (MMPs) have shown to result in MMPs-generated protein BDPs that are indicative of malignant growth in cancer, while in the field of neural injury, calpain-2 and caspase-3 proteases generate BDPs fragments that are indicative of different neural cell death mechanisms in different injury scenarios. Advanced proteomic techniques have shown a remarkable progress in identifying these BDPs experimentally. In this work, we present a bioinformatics-based prediction method that identifies protease-associated BDPs with high precision and efficiency. The method utilizes state-of-the-art sequence matching and alignment algorithms. It starts by locating consensus sequence occurrences and their variants in any set of protein substrates, generating all fragments resulting from cleavage. The complexity exists in space O(mn) as well as in O(Nmn) time, where N, m, and n are the number of protein sequences, length of the consensus sequence, and length per protein sequence, respectively. Finally, the proposed methodology is validated against βII-spectrin protein, a brain injury validated biomarker.

Degradomics discipline has been recently introduced to depict the application of an omics approach (genomics and proteomics etc.) to identify different proteases and their subsequent proteolytic substrates/degradome in a defined pathophysiological condition[1]. Recently, the use of bioinformatics tools as means for data mining has spanned different fields in cancer, neuroscience and biochemistry research[2,3]. Degradomics as a discipline has benefitted from data mining strategies as tools to predict degradome specific substrates *in silico*[4–7]. However, the application of bioinformatic tools on degradomics analysis requires different types of sequencing matching algorithms making it one of the challenging fields despite its potential beneficial outcomes mainly in clinical and diagnostic research. Knuth *et al.* developed an algorithm that only finds exact matches of a subsequence of size m in a sequence of size n in $O(m + n)$[8]. It is worth to know that other algorithms have identified sequence variants with comparable complexity, but not with the same fidelity. Lipman *et al.* devised a heuristic algorithm called FAST Protein (FASTP)[9]; it is based on alignment approach and is both rapid and sensitive in finding similarities between any amino acid subsequence and matching sequences in a database. Yet, it does not cover all regions, as it starts with an anchoring scheme that identifies identical regions using a replaceability matrix[9]. Similarly, Altschul *et al.* developed another heuristic algorithm BLAST, along with its variations; this algorithm supposedly supersedes FASTP in performance while retaining comparable sensitivity[10]. Nonetheless, it also uses seeds for basic anchoring as it identifies similar sequences to the query sequence by seeking segment pairs comprising a word pair of a given score.

[1]Department of Electrical and Computer Engineering, American University of Beirut, Riad El Solh, Beirut, Lebanon. [2]Department of Biochemistry and Molecular Genetics, Faculty of Medicine, American University of Beirut, Riad El Solh, Beirut, Lebanon. Correspondence and requests for materials should be addressed to Z.D. (email: zd03@aub.edu.lb) or F.K. (email: firasko@gmail.com)

In contrast, Ning *et al.* devised a sequence search and alignment algorithm based on the Sequence Search and Alignment by Hashing Algorithm (SSAHA); this method performs three to four times faster than FASTP or BLAST, as it handles searches in databases of gigabyte range[11]. However, it is associated with overhead as it pre-processes the sequences in the database by breaking them into consecutive k-tuples, and then uses a hash-table to store the position of each k-tuple occurrence. Ma *et al.* devised another search algorithm that works faster than BLAST, with both a modest memory usage and higher sensitivity, covering a wider seeding model[12]. Nevertheless, it is also based on heuristics, compromising accuracy to a certain extent. Kurtz *et al.* developed a suffix tree-based method for similarity sequence search, implemented with linked lists[13]. This method performs well, but is limited to exact searches and suffers from overhead due to large space requirements, with continual and necessary updates requirements to the linked lists. Lecroq developed an algorithm based on the Q-gram hashing; it is considered the fastest so far, especially on a small size alphabet, because it searches the sequence database using an efficient indexing technique[14]. Nonetheless, it is also limited to searching for exact matches. Needleman *et al.* created the first method for biological sequence comparison based on dynamic programming[15]. Even though this method is considered to be optimal, it is based on global alignment, which renders it more specific to sequences of comparable sizes.

Applications of degradomics studies have been witnessed in several diseases, such as brain injury and cancer[6,7,16,17]. In brain injury field, both calpain-2 and caspase-3 proteases generate signature protein markers that would theoretically be indicative of different types of neural cell injury mechanisms[16–18]. These signature markers are fragment proteins or BDPs, resulting from proteases-associated cleavages. Since they are differentiated by their sequence and molecular weight (Mwt) specificity, they are considered unique to each protease with a definitive signature Mwt characterized by a well-defined amino acid sequence. Degradomics-peptidomics profiling of blood plasma, for instance, showed high sensitivity to changes not evidenced by standard proteomics techniques, providing unique signatures of diagnostic utility[19].

In cancer research, Itoh *et al.* presented a review of the intense role of MMPs in cancer disease. Metalloproteases, MMP-2 and MMP-9, generate protein substrate fragments that are indicative of malignant growth[20]. Similarly, Fuhrman-Luck *et al.* used degradomics studies to identify kallikrein-related peptidase (KLK) substrates as biomarkers for cancer disease[21]. In Lopez-Otin *et al.* work, the local degradation of extra cellular matrix (ECM) forming the physical barrier for cell migration components is observed due to the activation of matrix MMPs[22]. MMPs, similar to other proteases (caspases, calpains and cathepsins) can truncate proteins at specific amino acid sequences[20,22].

On the other hand, degradomics studies have been noticeable in the genetic aspects of congenital heart disease (CHD) since they represent major causes of birth defects in newborns. A crucial gene in this context is the *TLL1* which encodes a metalloprotease. Upon activation, this metalloprotease truncates extracellular substrate proteins in the septum and the resulting BDPs represent putative markers of the disease[23]. Moreover, degradomics studies have contributed significantly to the field of neuroscience particularly in neural injury conditions. Glantz *et al.* discerned the molecular basis of protease-catalyzed proteolysis of αII-spectrin and βII-spectrin in the different injury scenarios, indicative of different neural injury techniques (both apoptotic and necrotic)[24].

The main potential of this work lies in its ability to predict fragments sequences computationally prior to wet experiments. This work extends on previous research in identifying brain injury specific BDPs, utilizing data of different potential substrate proteins, extracted from public databases[25]. The goal is to computationally search through the set of selected proteins for potential breakdown sites (consensus exact matches as well as variants), subject to fragmentation, and subsequently generate all possible cleaved fragments (BDPs). The work presents a dynamic programming solution based on modifications to the Smith-Waterman (SW) algorithm. Accordingly, the solution is based on local alignment and runs in time and space complexity O(mn) per protein sequence (m and n represent the sizes of the consensus sequence as well as the protein sequence, respectively). The method is applied to calpain-2 and caspase-3 proteases which are associated with the execution phase of both the apoptotic and necrotic cell death, and where the distinction between the two dominated types of cell death is crucial to better reveal the injury mechanisms.

This paper is organized as follows: We first describe the results of applying the computational method to the cleavage of βII-spectrin substrates, particularly in brain injury. It then describes the implicated proteases calpain-2 and caspase-3 as well as the distinction between the associated neural cell deaths (necrosis/apoptosis). After elaborating on the cleavage modes of calpain-2 and caspase-3, the results section presents the generated data for both the βII-spectrin protein and the mouse genome used in testing the algorithm and the corresponding output results. Afterwards, discussion of output is presented in the discussion section. Finally, the methods section defines the main problem and then describes the method and the algorithm involved in more details.

Results

Calpain-2 and Caspase-3 Proteases.

Both calpain-2 and caspase-3 are activated in different modes of neural cell death and; thus, it is essential to characterize their spatio-temporal activation as it is indicative of the injury mechanisms. Calpain-2 protease is activated in necrosis and apoptosis; it generates calpain-specific cleaved fragments. On the other hand, caspase-3 protease is only activated in neural apoptosis and generates caspase-specific cleaved fragments. Both of these proteases cleave their associated protein substrates at a cleavage site predefined within each consensus occurrence resulting in unique BDPs. Since the BDPs are differentiated by their protease-generated molecular weight (see Fig. 1), they are specific to each protease and represent molecular signatures[16,18,26].

In order to characterize the possible cleavage fragments of an already characterized protein, we assessed the brain injury biomarker βII-spectrin with its BDPs post calpain-2 and caspase-3 activation. We apply our proposed *Cleaved Fragments Prediction Algorithm for Calpain and Caspase* (*CFPA-CalpCasp*), that is specifically designed to assess calpain-2 and caspase-3 cleavage substrate sequences[25]. The algorithm is based on dynamic programming

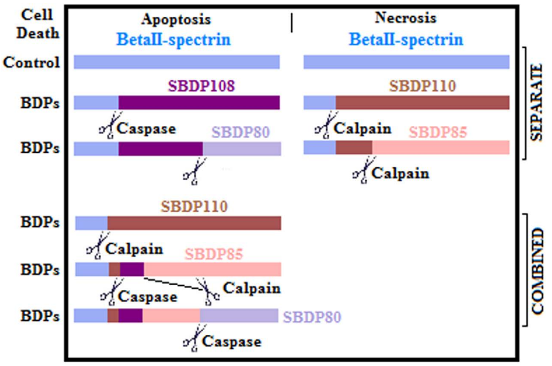


Figure 1. Schematic of necrosis and apoptosis cell death pathways. Figure shows necrosis type of neuronal cell death on the right, with calpain-2 specific fragments SBDP110 and SBDP85. On the left, the figure shows apoptosis neuronal cell death type, with caspase-3 specific fragments SBDP108 and SBDP80. The fragments are labeled with their approximate sizes. On the bottom, the figure shows the breakdown products by order of generation. First, calpain-2 cleaves βII-spectrin and generates SBDP110. Afterwards, and if both of caspase-3 and calpain-2 are activated, SBDP110 is cleaved by both proteases, generating SBDP108 and SBDP85. Lastly, caspase-3 cleaves SBDP85 and produces SBDP80 (apoptosis-specific)[16,32].

	Calpain	Caspase
Protease Class	Cysteine Protease	Cysteine Protease
Preferred Cleavage Site (*)	AspxxAsp*x	(Leu, Val, Ile)x*x
Common Substrates	αII-spectrin 280 kDa	αII-spectrin 280 kDa
	βII-spectrin 260 kDa	βII-spectrin 260 kDa
Fragments Produced by	SBDP110 kDa	SBDP108 kDa
βII-spectrin	SBDP85 kDa	SBDP80 kDa
Cell Death Involvement	Most forms of necrosis	Most forms of apoptosis
	Some forms of apoptosis	

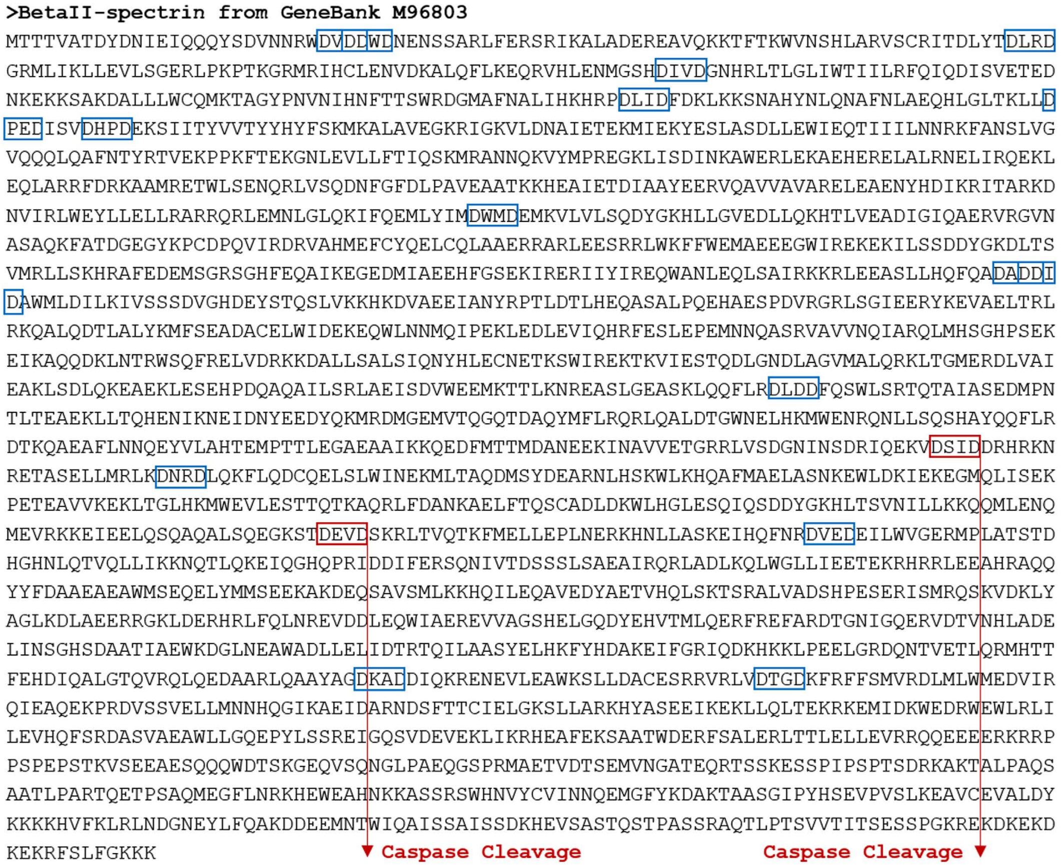
Table 1. Calpain-2 and Caspase-3 Cleavage Properties. The table shows the different properties specific to calpain and caspase. Most importantly, it shows the consensus patterns at which each of the proteases cleaves. For caspase protease, the consensus pattern is DXXD and the cleavage site is right after this pattern (D is Aspartic Acid and X is any amino acid). Calpain, on the other hand, cleaves after the consensus patterns LX, VX, or IX, where L is Leucine, V is Valine, and I is Isoleucine.

principles and is efficient in terms of both time and 'Space Complexity' (Run time) complexity. This algorithm is capable of performing local sequence alignment achieved via a scoring table; in addition, it is able to find consensus occurrences and variants of the consensus sequence after accounting for insert and delete operations[27].

The algorithm CFPA-CalpCasp proved its effectiveness after validating its results against experimental studies from previous works in the literature. This allows the utilization of the proposed computational methodology to guide and complement further experimental studies. Both of calpain-2 and caspase-3 proteases can lead to different cleaved fragments depending on whether the activation is combined or separate. The following three subsections illustrate the three different cases of when caspase-3 is activated separately, then when calpain-2 is activated separately, and finally when both of caspase-3 and calpain-2 are activated together (see summary in Table 1).

Caspase-3 Cleavage Mode (Apoptosis). Caspase-3, similar to calpain-2, is a cytosolic cysteine protease. However, caspase-3 differs from calpain-2 in the requirement for Ca^2. Furthermore, what distinguishes caspase-3 from other proteases is that it has a crucial role in apoptosis in many different cell types[27]. When caspase-3 is activated, it functions as a downstream mediator in apoptosis and exclusively generates βII-spectrin BDP specific fragment "SBDP108". Caspase-3 cleaves the substrate after finding both of Asp in the first position (P1 position) and Asp in the fourth (P4 position), whereas any amino acids can occupy the second position (P2 position) and the third position (P3 position), as indicated in Table 1[18].

Calpain-2 Cleavage Mode (Necrosis and Apoptosis). The association of calpain-mediated proteolysis to necrotic neuronal death has gained major research focus. This relation was revealed in ischemic and excitotoxic neural injury[17]. Calpain-2 cleaves the substrate after finding either of Val, Leu, or Ile residues in the target protein. It cleaves in the second (P2 position) after Val, Leu, or Ile amino acid is found in the first (P1 position). Accordingly, the P2 position in the target protein can be any residue (for example, Tyr, Gly, Arg) (as shown in Table 1)[18].



```
>BetaII-spectrin from GeneBank M96803
MTTTVATDYDNIEIQQQYSDVNNRWDVDDWDNENSSARLFERSRIKALADEREAVQKKTFTKWVNSHLARVSCRITDLYTDLRD
GRMLIKLLEVLSGERLPKPTKGRMRIHCLENVDKALQFLKEQRVHLENMGSHDIVDGNHRLTLGLIWTIILRFQIQDISVETED
NKEKKSAKDALLLWCQMKTAGYPNVNIHNFTTSWRDGMAFNALIHKHRPDLIDFDKLKKSNAHYNLQNAFNLAEQHLGLTKLLD
PEDISVDHPDEKSIITYVVTYYHYFSKMKALAVEGKRIGKVLDNAIETEKMIEKYESLASDLLEWIEQTIIILNNRKFANSLVG
VQQQLQAFNTYRTVEKPPKFTEKGNLEVLLFTIQSKMRANNQKVYMPREGKLISDINKAWERLEKAEHERELALRNELIRQEKL
EQLARRFDRKAAMRETWLSENQRLVSQDNFGFDLPAVEAATKKHEAIETDIAAYEERVQAVVAVARELEAENYHDIKRITARKD
NVIRLWEYLLELLRARRQRLEMNLGLQKIFQEMLYIMDWMDEMKVLVLSQDYGKHLLGVEDLLQKHTLVEADIGIQAERVRGVN
ASAQKFATDGEGYKPCDPQVIRDRVAHMEFCYQELCQLAAERRARLEESRRLWKFFWEMAEEEGWIREKEKILSSDDYGKDLTS
VMRLLSKHRAFEDEMSGRSGHFEQAIKEGEDMIAEEHFGSEKIRERIIYIREQWANLEQLSAIRKKRLEEASLLHQFQADADDI
DAWMLDILKIVSSSDVGHDEYSTQSLVKKHKDVAEEIANYRPTLDTLHEQASALPQEHAESPDVRGRLSGIEERYKEVAELTRL
RKQALQDTLALYKMFSEADACELWIDEKEQWLNNMQIPEKLEDLEVIQHRFESLEPEMNNQASRVAVVNQIARQLMHSGHPSEK
EIKAQQDKLNTRWSQFRELVDRKKDALLSALSIQNYHLECNETKSWIREKTKVIESTQDLGNDLAGVMALQRKLTGMERDLVAI
EAKLSDLQKEAEKLESEHPDQAQAILSRLAEISDVWEEMKTTLKNREASLGEASKLQQFLRDLDDFQSWLSRTQTAIASEDMPN
TLTEAEKLLTQHENIKNEIDNYEEDYQKMRDMGEMVTQGQTDAQYMFLRQRLQALDTGWNELHKMWENRQNLLSQSHAYQQFLR
DTKQAEAFLNNQEYVLAHTEMPTTLEGAEAAIKKQEDFMTTMDANEEKINAVVETGRRLVSDGNINSDRIQEKVDSIDDRHRKN
RETASELLMRLKDNRDLQKFLQDCQELSLWINEKMLTAQDMSYDEARNLHSKWLKHQAFMAELASNKEWLDKIEKEGMQLISEK
PETEAVVKEKLTGLHKMWEVLESTTQTKAQRLFDANKAELFTQSCADLDKWLHGLESQIQSDDYGKHLTSVNILLKKQQMLENQ
MEVRKKEIEELQSQAQALSQEGKSTDEVDSKRLTVQTKFMELLEPLNERKHNLLASKEIHQFNRDVEDEILWVGERMPLATSTD
HGHNLQTVQLLIKKNQTLQKEIQGHQPRIDDIFERSQNIVTDSSSLSAEAIRQRLADLKQLWGLLIEETEKRHRRLEEAHRAQQ
YYFDAAEAEAWMSEQELYMMSEEKAKDEQSAVSMLKKHQILEQAVEDYAETVHQLSKTSRALVADSHPESERISMRQSKVDKLY
AGLKDLAEERRGKLDERHRLFQLNREVDDLEQWIAEREVVAGSHELGQDYEHVTMLQERFREFARDTGNIGQERVDTVNHLADE
LINSGHSDAATIAEWKDGLNEAWADLLELIDTRTQILAASYELHKFYHDAKEIFGRIQDKHKKLPEELGRDQNTVETLQRMHTT
FEHDIQALGTQVRQLQEDAARLQAAYAGDKADDIQKRENEVLEAWKSLLDACESRRVRLVDTGDKFRFFSMVRDLMLWMEDVIR
QIEAQEKPRDVSSVELLMNNHQGIKAEIDARNDSFTTCIELGKSLLARKHYASEEIKEKLLQLTEKRKEMIDKWEDRWEWLRLI
LEVHQFSRDASVAEAWLLGQEPYLSSREIGQSVDEVEKLIKRHEAFEKSAATWDERFSALERLTTLELLEVRRQQEEEERKRRP
PSPEPSTKVSEEAESQQQWDTSKGEQVSQNGLPAEQGSPRMAETVDTSEMVNGATEQRTSSKESSPIPSPTSDRKAKTALPAQS
AATLPARTQETPSAQMEGFLNRKHEWEAHNKKASSRSWHNVYCVINNQEMGFYKDAKTAASGIPYHSEVPVSLKEAVCEVALDY
KKKKHVFKLRLNDGNEYLFQAKDDEEMNTWIQAISSAISSDKHEVSASTQSTPASSRAQTLPTSVVTITSESSPGKREKDKEKD
KEKRFSLFGKKK
```

Figure 2. Cleavage sites of βII-spectrin by caspase-3. The figure shows all the consensus subsequences predicted by the algorithm, surrounded in boxes, and obeying the amino acid sequence DXXD (D stands for Asp amino acid and X can be any amino acid). In particular, the red boxes represent the consensus occurrences validated experimentally. In addition, the figure shows the cleavage site where caspase-3 cleaves.

Combined Cleavages of Calpain-2 and Caspase-3. Calpain-2 protease is usually activated before caspase-3. This temporal profile allows both proteases to cleave one protein substrate at separate cleavage sites without interference. It is possible though, in a random and infrequent instance, for calpain-2 and caspase-3 to be activated concurrently. Some injury models reveal the synchronized activation of both proteases like the *in vivo* model of traumatic brain injury (TBI), which affects different areas of the brain. In addition, other neuronal injury mechanisms, demonstrating the activation of both proteases, include NMDA, kainate, and glucose–oxygen-deprivation cerebrocortical neurons[28–30].

Input Data. The algorithm needs to be validated with real data to verify its accuracy and effectiveness. Thus, the substrate βII-spectrin is used for input data, as shown in Supplementary Fig. 1. In addition, the mouse genome is also used as input data to test the efficiency of the algorithm on a large dataset[31].

Output of βII-spectrin Cleavage by Caspase-3. The pattern DXXD↑X corresponds to the consensus sequence for caspase-3 protease, where X represents any one amino acid from the twenty primary amino acids, symbol ↑ represents the site of cleavage, and D represents Aspartic acid (Asp) amino acid. All different combinations of the above pattern correspond to 400 expected instances in total. The partial amino acid subsequences, presented in Fig. 2, highlight caspase-3 consensus occurrences showing two *hits* in red that are validated experimentally[16,32]. Figure 2 also shows caspase-3 protease cleavage mode in cleaving an input protein sequence substrate[25]. The results of applying CFPA-CalpCasp algorithm on βII-spectrin input protein sequence and caspase-3 protease are shown in Table 2. The table indicates all the consensus occurrences (*hits*) detected by caspase-3 protease for cleavage, including the cleavage site corresponding to each consensus occurrence. The start and end positions of each consensus occurrence within the given input protein sequence are also indicated. Furthermore, the table shows all the fragments generated from cleaving the input protein sequence, at the detected consensus occurrences and cleavage sites, including their start and end positions.

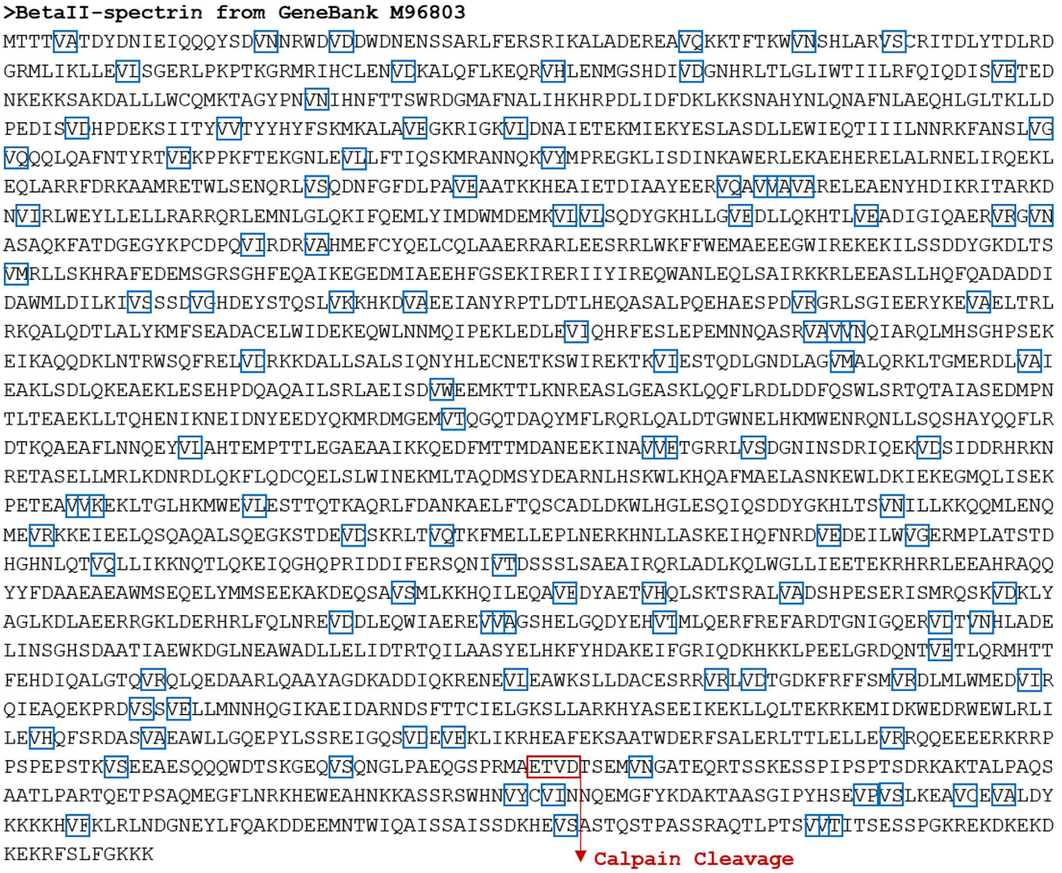
From the detected hits, we list the specific subsequences 'DEVD' and 'DSID', which are validated against the experimentally generated fragments[16,32]. Motif 'DSID' starts at position 1251 (or P1251) and ends at position 1254 (or P1254) within βII-spectrin input protein sequence. The corresponding cleaved subsequence fragments are 'MTTT…DSID', which starts at position 1 (P1) and ends at position 1254 (P1254), and 'DRHR… GKKK', which starts at position 1255 (P1255) and ends at position 2364 (P2364). On the other hand, the start position of motif

		AA Sequence	Start	End			AA Sequence	Start	End
1	Consensus	DVDD	26	29	10	Consensus	DDID	754	757
	Fragment	MTTT.... DVDD	1	29		Fragment	MTTT.... DDID	1	757
	Fragment	WDNE.... GKKK	30	2364		Fragment	AWML.... GKKK	758	2364
2	Consensus	DDWD	28	31	11	Consensus	DLDD	1070	1073
	Fragment	MTTT.... DDWD	1	31		Fragment	MTTT.... DLDD	1	1073
	Fragment	NENS.... GKKK	32	2364		Fragment	FQSW.... GKKK	1074	2364
3	Consensus	DLRD	81	84	12	Consensus	DSID	1251	1254
	Fragment	MTTT.... DLRD	1	84		Fragment	MTTT.... DSID	1	1254
	Fragment	GRML.... GKKK	85	2364		Fragment	DRHR.... GKKK	1255	2364
4	Consensus	DIVD	137	140	13	Consensus	DNRD	1273	1276
	Fragment	MTTT.... DIVD	1	140		Fragment	MTTT.... DNRD	1	1276
	Fragment	GNHR.... GKKK	141	2364		Fragment	LQKF.... GKKK	1277	2364
5	Consensus	DLID	218	221	14	Consensus	DEVD	1454	1457
	Fragment	MTTT.... DLID	1	221		Fragment	MTTT...DEVD	1	1457
	Fragment	FDKL.... GKKK	222	2364		Fragment	SKRL...GKKK	1458	2364
6	Consensus	DPED	252	255	15	Consensus	DVED	1493	1496
	Fragment	MTTT.... DPED	1	255		Fragment	MTTT.... DVED	1	1496
	Fragment	ISVD.... GKKK	256	2364		Fragment	EILW.... GKKK	1497	2364
7	Consensus	DHPD	259	262	16	Consensus	DKAD	1877	1880
	Fragment	MTTT.... DHPD	1	262		Fragment	MTTT.... DKAD	1	1880
	Fragment	EKSI.... GKKK	263	2364		Fragment	DIQK.... GKKK	1881	2364
8	Consensus	DWMD	542	545	17	Consensus	DTGD	1909	1912
	Fragment	MTTT.... DWMD	1	545		Fragment	MTTT.... DTGD	1	1912
	Fragment	EMKV.... GKKK	546	2364		Fragment	KFRF.... GKKK	1913	2364
9	Consensus	DADD	752						
	Fragment	MTTT.... DADD	1						
	Fragment	IDAW.... GKKK	756						

Table 2. CFPA-CalpCasp Generated Data on M96803 (βII-spectrin Protein Sequence) by Caspase-3. Table 2 shows the consensus occurrences detected by CFPA-CalpCasp algorithm, specific to caspase protease. Accordingly, it shows the detected pattern obeying DXXD (D is Aspartic Acid and X is any amino acid) and its start and end positions within the βII-spectrin protein sequence. Afterwards, the algorithm generates all the fragments resulting from caspase cleavage based on the corresponding consensus occurrence. The fragments are listed with their start and end positions within the βII-spectrin protein sequence. Red highlights show the consensus occurrences and the corresponding cleavages that are validated experimentally.

'DEVD', within βII-spectrin input protein sequence, is at P1454, and its end position is at P1457. The corresponding cleaved subsequence fragments are 'MTTT…DEVD', which starts at position 1 (P1) and ends at position 1457 (P1457), and 'SKRL…GKKK', which starts at position 1458 (P1458) and ends at position 2364 (P2364).

Output of βII-spectrin Cleavage by Calpain-2. The patterns of LX↑X, VX↑X, and IX↑X correspond to the consensus sequences of calpain-2 protease, where X represents any amino acid from the twenty primary amino acids, while symbol ↑ represents the cleavage site, and (L, V, I) triplet maps to (Leu, Val, Ile) amino acid triplet; respectively. All different combinations of the above patterns correspond to 60 expected instances in total. The partial amino acid subsequences, pictured in Fig. 3, highlight calpain-2 consensus occurrences, showing one *hit* in red that is validated experimentally[16,32]. Figure 3 also shows calpain-2 protease cleavage mode in cleaving an input protein sequence substrate[25]. The results of applying the proposed algorithm on βII-spectrin input



```
>BetaII-spectrin from GeneBank M96803
MTTTVATDYDNIEIQQQYSDVNNRWDVDDWDNENSSARLFERSRIKALADEREAVQKKTFTKWVNSHLARVSCRITDLYTDLRD
GRMLIKLLEVLSGERLPKPTKGRMRIHCLENVDKALQFLKEQRVHLENMGSHDIVDGNHRLTLGLIWTIILRFQIQDISVETED
NKEKKSAKDALLLWCQMKTAGYPNVNIHNFTTSWRDGMAFNALIHKHRPDLIDFDKLKKSNAHYNLQNAFNLAEQHLGLTKLLD
PEDISVDHPDEKSIITYVVTYYHYFSKMKALAVEGKRIGKVLDNAIETEKMIEKYESLASDLLEWIEQTIIILNNRKFANSLVG
VQQQLQAFNTYRTVEKPPKFTEKGNLEVLLFTIQSKMRANNQKVYMPREGKLISDINKAWERLEKAEHERELALRNELIRQEKL
EQLARRFDRKAAMRETWLSENQRLVSQDNFGFDLPAVEAATKKHEAIETDIAAYEERVQAVVAVARELEAENYHDIKRITARKD
NVIRLWEYLLELLRARRQRLEMNLGLQKIFQEMLYIMDWMDEMKVIVLSQDYGKHLLGVEDLLQKHTLVEADIGIQAERVRGVN
ASAQKFATDGEGYKPCDPQVIRDRVAHMEFCYQELCQLAAERRARLEESRRLWKFFWEMAEEEGWIREKEKILSSDDYGKDLTS
VMRLLSKHRAFEDEMSGRSGHFEQAIKEGEDMIAEEHFGSEKIRERIIYIREQWANLEQLSAIRKKRLEEASLLHQFQADADDI
DAWMLDILKIVSSSDVGHDEYSTQSLVKKHKDVAEEIANYRPTLDTLHEQASALPQEHAESPDVRGRLSGIEERYKEVAELTRL
RKQALQDTLALYKMFSEADACELWIDEKEQWLNNMQIPEKLEDLEVIQHRFESLEPEMNNQASRVAVVNQIARQLMHSGHPSEK
EIKAQQDKLNTRWSQFRELVDRKKDALLSALSIQNYHLECNETKSWIREKTKVIESTQDLGNDLAGVMALQRKLTGMERDLVAI
EAKLSDLQKEAEKLESEHPDQAQAILSRLAEISDVWEEMKTTLKNREASLGEASKLQQFLRDLDDFQSWLSRTQTAIASEDMPN
TLTEAEKLLTQHENIKNEIDNYEEDYQKMRDMGEMVTQGQTDAQYMFLRQRLQALDTGWNELHKMWENRQNLLSQSHAYQQFLR
DTKQAEAFLNNQEYVLAHTEMPTTLEGAEAAIKKQEDFMTTMDANEEKINAVVETGRRLVSDGNINSDRIQEKVDSIDDRHRKN
RETASELLMRLKDNRDLQKFLQDCQELSLWINEKMLTAQDMSYDEARNLHSKWLKHQAFMAELASNKEWLDKIEKEGMQLISEK
PETEAVVKEKLTGLHKMWEVLESTTQTKAQRLFDANKAELFTQSCADLDKWLHGLESQIQSDDYGKHLTSVNILLKKQQMLENQ
MEVRKKEIEELQSQAQALSQEGKSTDEVDSKRLTVQTKFMELLEPLNERKHNLLASKEIHQFNRDVEDEILWVGERMPLATSTD
HGHNLQTVQLLIKKNQTLQKEIQGHQPRIDDIFERSQNIVTDSSSLSAEAIRQRLADLKQLWGLLIEETEKRHRRLEEAHRAQQ
YYFDAAEAEAWMSEQELYMMSEEKAKDEQSAVSMLKKHQILEQAVEDYAETVHQLSKTSRALVADSHPESERISMRQSKVDKLY
AGLKDLAEERRGKLDERHRLFQLNREVDDLEQWIAEREVVAGSHELGQDYEHVTMLQERFREFARDTGNIGQERVDTVNHLADE
LINSGHSDAATIAEWKDGLNEAWADLLELIDTRTQILAASYELHKFYHDAKEIFGRIQDKHKKLPEELGRDQNTVETLQRMHTT
FEHDIQALGTQVRQLQEDAARLQAAYAGDKADDIQKRENEVIEAWKSLLDACESRRVRLVDTGDKFRFFSMVRDLMLWMEDVIR
QIEAQEKPRDVSSVELLMNNHQGIKAEIDARNDSFTTCIELGKSLLARKHYASEEIKEKLLQLTEKRKEMIDKWEDRWEWLRLI
LEVHQFSRDASVAEAWLLGQEPYLSSREIGQSVDEVEKLIKRHEAFEKSAATWDERFSALERLTTLELLEVRRQQEEEERKRRP
PSPEPSTKVSEEAESQQQWDTSKGEQVSQNGLPAEQGSPRMAETVDTSEMVNGATEQRTSSKESSPIPSPTSDRKAKTALPAQS
AATLPARTQETPSAQMEGFLNRKHEWEAHNKKASSRSWHNVYCVINNQEMGFYKDAKTAASGIPYHSEVPVSLKEAVCEVALDY
KKKKHVEKLRLNDGNEYLFQAKDDEEMNTWIQAISSAISSDKHEVSASTQSTPASSRAQTLPTSVVTITSESSPGKREKDKEKD
KEKRFSLFGKKK
```

Figure 3. Cleavage sites of βII-spectrin by calpain-2. The figure shows all the consensus subsequences predicted by the algorithm, surrounded in boxes, and conforming with the amino acid sequence VX (V stands for Val amino acid and X can be any amino acid). In particular, the red box represents the consensus occurrence validated experimentally. In addition, the figure shows the cleavage site where calpain-2 cleaves.

protein sequence and calpain-2 protease are shown partially in Table 3 (see Supplementary Table 1). The table lists all the hits detected by calpain-2 protease for cleavage, including the cleavage site corresponding to each consensus occurrence, and the start and end position of each consensus occurrence within the input protein sequence. Furthermore, the table shows all the fragments generated from cleaving the input protein sequence, at the detected consensus occurrences and cleavage sites, including their start and end positions.

From the detected hits, we list the specific subsequence 'ETVD', which is validated against the experimentally generated fragments[16,32]. Motif 'ETVD' starts at position 2143 (or P2143) and ends at position 2146 (or P2146) within βII-spectrin input protein sequence. The corresponding generated sequence fragments are 'MTTT… ETVD' and 'TSEM… GKKK'. The first fragment extends from position 1 (P1) through position 2146 (P2146), and the second one extends from position 2147 (P2147) through position 2364 (P2364). The other detected occurrences, such as 'VH', 'VA', 'IK', and 'LM' (see Fig. 3), did not appear in experimental results, and the reason could be linked to the rapid pace of the cleavage transitions. Particularly, the end-to-end cleavage sites may obscure the digestion of the detected sequence occurrences in a simultaneous manner, and thus may end up in undetectable occurrences by experimental techniques.

Simultaneous activation of both proteases (calpain-2 and caspase-3) has also been generated computationally and provides similar outcomes and insights to the separate activation of each protease. For experimental validation, two possibilities can arise in such a case: 1) one protease inhibits the cleavage of the other protease, or 2) one protease cleaves within the input sequence cleaved by the other protease.

Output of Mouse Proteome Cleavage by Caspase-3.

In order to assess the efficiency of the algorithm on a large dataset, the proposed algorithm is applied on the whole mouse proteome. Supplementary Table 2 shows all the consensus occurrences that result from the cleavage of the mouse proteome input protein sequences by caspase-3. The consensus occurrences appear per each protein sequence, including their start and end positions. Red highlights in Supplementary Table 2 cover the same consensus occurrence that appears multiple times within the same input protein sequence, including all the different start and end positions. On the other hand, blue highlights cover the consensus occurrences that overlap within one protein sequence. All consensus occurrences reflect the DXXD consensus pattern, where D represents Aspartic Acid and X represents any amino acid.

Table 4 depicts a detailed case of an input protein sequence having multiple occurrences and consensus overlaps. Supplementary Table 3 shows the corresponding cleaved fragments with their start and end positions.

		AA Sequence	Start	End			AA Sequence	Start	End
43	Consensus	PDVR	820	821	98	Consensus	ETVD	2145	2146
	Fragment	MTTT.... PDVR	1	821		Fragment	MTTT...ETVD	1	2146
	Fragment	GRLS.... GKKK	822	2364		Fragment	TSEM... GKKK	2147	2364
44	Consensus	KEVA	834	835	99	Consensus	EMVN	2151	2152
	Fragment	MTTT.... KEVA	1	835		Fragment	MTTT.... EMVN	1	2152
	Fragment	ELTR.... GKKK	836	2364		Fragment	GATE.... GKKK	2153	2364
45	Consensus	LEVI	886	887	100	Consensus	HNVY	2225	2226
	Fragment	MTTT.... LEVI	1	887		Fragment	MTTT.... HNVY	1	2226
	Fragment	QHRF.... GKKK	888	2364		Fragment	CVIN.... GKKK	2227	2364

Table 3. Few Records of CFPA-CalpCasp Generated Data on M96803 (βII-spectrin Protein Sequence) by Calpain-2 (see Supplementary Table 1 for all Output Records). Table 3 shows the consensus occurrences detected by CFPA-CalpCasp algorithm, specific to calpain protease. Accordingly, it shows *few* detected patterns obeying VX (V is Valine and X is any amino acid) and its start and end positions within the βII-spectrin protein sequence. Afterwards, the algorithm generates all fragments resulting from calpain cleavage based on the corresponding consensus occurrence. The fragments are listed with their start and end positions within the βII-spectrin protein sequence. Red highlights show the consensus occurrence and the corresponding cleavage that are validated experimentally.

Nevertheless, Supplementary Table 3 lists all possible cleavage combinations of the case; Supplementary Table 3 is simplified by showing the beginning and ending of each fragment subsequence.

Discussion

Computational prediction of biomarkers is becoming a priority for biologists, as it conserves both time and cost that would have been otherwise spent on experiments, in order to probe for biomarkers. The developed algorithm for cleaved fragments prediction (CFPA-CalpCasp) is based on Smith-Waterman algorithm and detects local subsequence similarities in a set of protein sequences[25]. Accordingly, alignments with deletions and insertions are pruned. Then, for every acceptable alignment, the protein sequence is cleaved at the predefined cleavage site within the consensus occurrence, and results in cleaved fragments identified by the algorithm. The consensus occurrence variants are built within the consensus pattern for calpain and caspase, such as in subsequence DXXD, where D is fixed for Asp, but X can be any amino acid.

To assess the effectiveness of the algorithm in identifying consensus subsequences, proteolysis, and fragment breakdowns generation, it is applied to βII-spectrin substrates. The corresponding results are validated with experimental data from the literature, demonstrating the accuracy of the algorithm. In addition, the algorithm proved its efficiency in performance by detecting all the consensus variants, cleaving them at similar cleavage site, and generating all potential BDPs with relatively low time and space complexity.

Furthermore, for a better efficiency assessment, the algorithm is applied on a large dataset (mouse genome) comprising ~30 k protein sequences[31]. Once more, the algorithm proved its performance efficiency by detecting the consensus variants, cleaving them at the cleavage site, and generating the resulting breakdown products (refer to Algorithm in Supplementary Fig. 2). Moreover, the results demonstrated the functionality of all "*different and possible*" types of cleavage combinations, in addition to the functionality of "*overlapping*" consensus occurrences (refer to Supplementary Table 2).

The generated data and results of this research can help guide future experiments. To make the method accessible to the scientific community, a web-based front end tool will be developed for online access by users; the application will have a database backend which will store protein substrates, relevant consensus subsequences, and the generated breakdown products (BDPs or biomarkers). The front end web interface will comprise different types of functionality. Figure 4 below presents a preliminary mockup interface of the web tool. The major functionality is that scientists will be able to select a protease, a substrate protein, and a protease cleavage mode from drop down menus; after selection of input data from the web form, they will execute the algorithm to obtain all the fragments generated upon proteases cleavage of substrate proteins. The corresponding output of biomarkers will then be presented in a list that scientists can scroll through and download for further post-processing.

Another functionality we are building into the web tool is the ability to click on a specific biomarker and output all the corresponding properties, by linking to public databases. Such information is crucial for experimentalists as it allows them to identify whether a specific biomarker has an existing antibody, instead of designing a new one. Furthermore, the above strategies can be applied to other disciplines utilizing degradomics as means for biomarker identification.

Input Protein Sequence (Seq. # 194)
MLQDSITGIVNSFNLFFPSTMSRPTLMPTCVAFCSILFLTLATGCQAFPKVERRETAQEYAEKEQSQKMNTD DQENISFAPKYMLQQMSSEAPMVLSEGPSEIPLIKVFSVNKESHLPGAGLLHPTSPGVYSSSEPVVSASEQ EPGPSLLERMSSEHSLSKVMLTVAVSSPASLNPDQEGPYNSLSTQPIVAAVTDVTHGSLDYLDNQLFAAKS QEAVSLGNSPSSSINTKEPEIIKADAAMGTTVVPGVDSTGDMEPDRERPSEMAADDGQSTTTKYLVTIPNNFL TTEPTAGSILGDAKVTVSVSTAGPVSSIFNEEWDDTKFESISRGRPPEPGDNAETQMRTKPPHGTYESFEGT EESPSSTAVLKVAPGHLGGEPALGTALVTALGDERSPVLTHQISFTPMSLAEDPEVSTMKLFPSAGGFRASTQG DRTQLSSETAFSTSQYESVPQQEAGNVLKDITQERKMATQAMNTTSPVVTQEHMATIEVPRGSGEPEEGMP SLSPVPAEVADAELSRRGESLATPASTTVVPLSLKLTSSMEDLMDTITGPSEEFIPVLGSPMAPPAMTVEAPTIS SALPSEGRTSPSISRPNTAASYGLEQLESEEVEDDEDEEDEEDEEEEEEDEEDEEDEEDKETDSLYKDFDGDTEPPG FTLPGITSQEPDIRSGSMDLLEVATYQVPETIEWEQQNQGLVRSWMEKLKDKAGYMSGMLVPVGVGIAGALF ILGALYSIKVMNRRRRNGFKRHKRKQREFNSMQDRVMLLADSSEDEF

Sequence Number	Consensus Occurrence #	Consensus	Start	End
194	1	DEED	616	619
	2	DEED	619	622
	3	DEED	629	632
	4	DEED	632	635
	5	DEED	635	638

Table 4. Input Protein Sequence with Multiple Consensus Occurrences including Overlaps. Table 4 shows a sample protein sequence from the whole mouse genome. This specific protein sequence is a sound case showing two functionalities implemented within the algorithm. The first is the multiple occurrences of the same consensus sequence (DEED) within the same protein sequence. The second is the overlapping consensus occurrences illustrated by their start and end positions (first DEED ends at 619 and next DEED starts at 619).

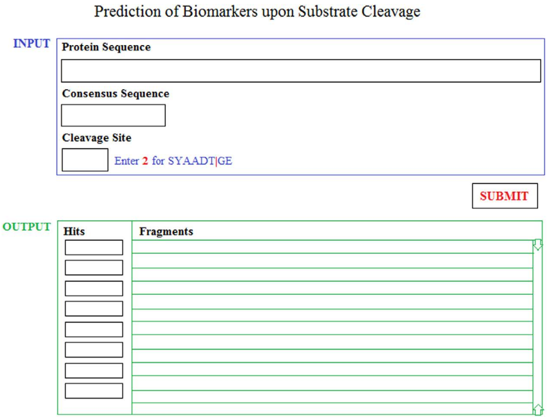


Figure 4. Preliminary mockup interface of the web tool. The figure shows a preliminary design for the web interface that will be developed to provide researchers with access to the proposed algorithm.

Methods

Problem Definition.

The problem is to locate all consensus occurrences of a consensus subsequence in a set of protein sequences. Once the consensus subsequence is detected on the protein sequence, the protease enzyme can then cleave a protein substrate at the predefined *cleavage site* within the consensus subsequence. This results in the formation of fragment subsequences or BDPs, signifying disease biomarkers. The output is expected to show all occurrences (*hits*) of the consensus sequence among all input protein sequences, in addition to the corresponding cleaved fragments. The hits correspond to exact matches of the consensus model; this model also contains variants within itself (shown in Table 1 as a combination of fixed and variable amino acids).

We elaborate on the CFPA-CalpCasp algorithm in the next section. The space (Run time) complexity of CFPA-CalpCasp is O(mn) and the time (computational) complexity is O(NN'mn), where N, N', n, and m are the total number of input protein sequences, total number of consensus sequences, size per protein sequence, and size per consensus sequence, respectively.

Computational Method.

The goal of the developed method CFPA-CalpCasp is to detect all consensus occurrences (and variants) of a specifically known consensus subsequence - with a specifically known cleavage site – in a set of input protein sequences. In addition, this method is capable to identify generated cleaved fragments upon cleavage of the input sequences[25]. Once the consensus subsequence is found (or matched) in an input protein sequence substrate, the activated protease enzyme cleaves the input sequence at the cleavage site - predefined within the consensus subsequence - resulting in fragment subsequences or breakdowns (BDPs).

Due to the specificity of calpain-2 and caspase-3 cleavage modes, CFPA-CalpCasp looks for an exact match of every stored consensus subsequence. The consensus variants are generated by fixing certain amino acids while varying others within each stored consensus subsequence. Accordingly, the *cleavage site* becomes right after the consensus *hit*. The algorithm embeds a modification version of Smith-Waterman algorithm; it performs local sequence alignments which allow to identify all local regions within each input protein sequence that match a particular consensus sequence[33]. The alignments are based on dynamic programming technique which constructs a scoring table, but then removes all consensus occurrences with INserts or DELetes (INDELs). To process N input protein sequences, the algorithm executes N times.

If multiple occurrences of any consensus subsequence are found in one protein sequence, the proteases might cleave at the cleavage site of each consensus occurrence. Consequently, all different combinations of potential cleavages are possible because each combination is actually a possible cleavage incidence by nature. Accordingly, the fragment generation module, within the developed algorithm, generates different scenarios of output fragments based on the different combinations of consensus occurrences. The following represents an illustration of the different output fragments generated by the algorithm upon detecting two consensus occurrences in one protein sequence:

(a) The algorithm generates the output fragments after cleaving the input protein sequence at the cleavage site of the first occurrence, resulting in fragment 1 and fragment 2.
(b) The algorithm generates the output fragments after cleaving the input protein sequence at the cleavage site of the second occurrence, resulting in fragment 3 and fragment 4.
(c) The algorithm generates the output fragments after cleaving the input protein sequence at the cleavage sites of the first and second occurrences, resulting in fragment 1, fragment 4, and fragment 5 (which is located between the two cleavage sites).

The last case c) results in many short fragments, compared to a few long ones from the first two cases a) and b).

Due to space limitations, we are not presenting the fragments generated from all combinations. Tables 2 and 3 show exact consensus matches and the corresponding cleaved fragments by caspase-3 and calpain-2 on βII-spectrin substrate respectively. Table 4 and Supplementary Table 3 show caspase-3 exact consensus matches from the mouse genome and the corresponding cleaved fragments. Moreover, the data shows the case of a single consensus with multiple occurrences within a single input protein sequence, including different occurrences that overlap (refer to Supplementary Table 2).

References

1. McQuibban, G. A. *et al.* Inflammation dampened by gelatinase A cleavage of monocyte chemoattractant protein-3. *Science* (*New York, N.Y.*) **289,** 1202–1206 (2000).
2. Alawieh, A. *et al.* Bioinformatics approach to understanding interacting pathways in neuropsychiatric disorders. *Methods Mol Biol* **1168,** 157–172, doi: 10.1007/978-1-4939-0847-9_9 (2014).
3. Godovac-Zimmermann, J. The 9th Siena meeting: from genome to proteome: open innovations. *Expert Rev Proteomics* **9,** 591–594, doi: 10.1586/epr.12.56 (2012).
4. Huesgen, P. F. & Overall, C. M. N- and C-terminal degradomics: new approaches to reveal biological roles for plant proteases from substrate identification. *Physiol Plant* **145,** 5–17, doi: 10.1111/j.1399-3054.2011.01536.x (2012).
5. Doucet, A. & Overall, C. M. Protease proteomics: revealing protease *in vivo* functions using systems biology approaches. *Mol Aspects Med* **29,** 339–358, doi: 10.1016/j.mam.2008.04.003 (2008).
6. Kobeissy, F. H., Sadasivan, S., Liu, J., Gold, M. S. & Wang, K. K. Psychiatric research: psychoproteomics, degradomics and systems biology. *Expert Rev Proteomics* **5,** 293–314, doi: 10.1586/14789450.5.2.293 (2008).
7. Overall, C. M. & Dean, R. A. Degradomics: systems biology of the protease web. Pleiotropic roles of MMPs in cancer. *Cancer Metastasis Rev* **25,** 69–75, doi: 10.1007/s10555-006-7890-0 (2006).
8. Knuth, D., Morris, J. & Pratt, V. Fast Pattern Matching in Strings. *SIAM Journal on Computing* **6** (1977).
9. Lipman, D. J. & Pearson, W. R. Rapid and sensitive protein similarity searches. *Science* (*New York, N.Y.*) **227,** 1435–1441 (1985).
10. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410, doi: 10.1016/S0022-2836(05)80360-2 (1990).
11. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res* **11,** 1725–1729, doi: 10.1101/gr.194201 (2001).
12. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18,** 440–445 (2002).
13. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5,** R12, doi: 10.1186/gb-2004-5-2-r12 (2004).
14. Lecroq, T. Fast exact string matching algorithms. *Information Processing Letters* **102,** 229–235, doi: 10.1016/j.ipl.2007.01.002 (2007).
15. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48,** 443–453, doi: 0022-2836(70)90057-4 (1970).
16. Wang, K. K. *et al.* Simultaneous degradation of alphaII- and betaII-spectrin by caspase 3 (CPP32) in apoptotic cells. *J Biol Chem* **273,** 22490–22497 (1998).
17. Yuen, P.-W. & Wang, K. W. Calpain inhibitors: Novel neuroprotectants and potential anticataract agents. *Drugs of the Future* **23,** doi: 10.1358/dof.1998.023.07.858362 (1998).
18. Wang, K. K. Calpain and caspase: can you tell the difference? by kevin K.W. Wang, Vol. 23, pp. 20–26. *Trends Neurosci* **23,** 59, doi: S0166-2236(99)01536-2 (2000).
19. Shen, Y. *et al.* Blood peptidome-degradome profile of breast cancer. *PLoS One* **5,** e13133, doi: 10.1371/journal.pone.0013133 (2010).
20. Itoh, Y. & Nagase, H. Matrix metalloproteinases in cancer. *Essays Biochem* **38,** 21–36 (2002).
21. Fuhrman-Luck, R. A. *et al.* Proteomic and other analyses to determine the functional consequences of deregulated kallikrein-related peptidase (KLK) expression in prostate and ovarian cancer. *Proteomics Clin Appl* **8,** 403–415, doi: 10.1002/prca.201300098 (2014).
22. Lopez-Otin, C. & Overall, C. M. Protease degradomics: a new challenge for proteomics. *Nat Rev Mol Cell Biol* **3,** 509–519, doi: 10.1038/nrm858 (2002).
23. Clark, T. G. *et al.* The mammalian Tolloid-like 1 gene, Tll1, is necessary for normal septation and positioning of the heart. *Development* **126,** 2631–2642 (1999).
24. Glantz, S. B. *et al.* Sequential degradation of alphaII and betaII spectrin by calpain in glutamate or maitotoxin-stimulated cells. *Biochemistry* **46,** 502–513, doi: 10.1021/bi061504y (2007).

25. El-Assaad, A., Dawy, Z., Nemer, G. & Kobeissy, F. Cleaved Fragments Prediction Algorithm (CFPA) application to calpain and caspase in apoptosis and necrotic cell death. *2015 IEEE International Conference on Electro/Information Technology* (*EIT*), 210–215, doi: 10.1109/EIT.2015.7293342 (2015).
26. Kobeissy, F. H. *et al.* Degradation of betaII-Spectrin Protein by Calpain-2 and Caspase-3 Under Neurotoxic and Traumatic Brain Injury Conditions. *Mol Neurobiol* **52,** 696–709, doi: 10.1007/s12035-014-8898-z (2015).
27. Nicholson, D. W. & Thornberry, N. A. Caspases: killer proteases. *Trends Biochem Sci* **22,** 299–306, doi: S0968-0004(97)01085-2 (1997).
28. Nath, R., Probert, A. Jr., McGinnis, K. M. & Wang, K. K. Evidence for activation of caspase-3-like protease in excitotoxin- and hypoxia/hypoglycemia-injured neurons. *J Neurochem* **71,** 186–195 (1998).
29. Pike, B. R. *et al.* Regional calpain and caspase-3 proteolysis of alpha-spectrin after traumatic brain injury. *Neuroreport* **9,** 2437–2442 (1998).
30. Pike, B. R. *et al.* Temporal relationships between de novo protein synthesis, calpain and caspase 3-like protease activation, and DNA fragmentation during apoptosis in septo-hippocampal cultures. *J Neurosci Res* **52,** 505–520, doi: 10.1002/(SICI)1097-4547(19980601)52:5<505::AID-JNR3>3.0.CO;2-G (1998).
31. Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E. & Blake, J. A. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* **36,** D724–728, doi: 10.1093/nar/gkm961 (2008).
32. Kobeissy, F. H. *et al.* Degradation of betaII-Spectrin Protein by Calpain-2 and Caspase-3 Under Neurotoxic and Traumatic Brain Injury Conditions. *Mol Neurobiol* **52,** 696–709, doi: 10.1007/s12035-014-8898-z (2014).
33. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol* **147,** 195–197, doi: 0022-2836(81)90087-5 (1981).

Acknowledgements

This work was undertaken at the American University of Beirut (AUB). The research was supported by a Biomedical Research Program Project within AUBMC [Grant title: The Identification of the "Compensatory Septal Proteome" (CSP) and Tll1 Substrates Involved in CARDIAC SEPTATION: Potential for Cardiac Defect Biomarkers; Lead PI: Georges Nemer] and by a grant from the National Council for Scientific Research (CNRS) in Lebanon [Grant title: Computational Modeling of DNA Repair Mechanism (MMR) for Understanding Defects by MMR Leading to Congenital Heart Diseases (CHD); Lead PI: Zaher Dawy]. We thank Edwyn Assaf and Alaa Kassir for their help in English Editing.

Author Contributions

A.E. designed, developed, and tested the algorithm on real data (Mouse Genome). A.E. prepared preliminary data, cleaned and analyzed the real data, interpreted the results, and wrote the manuscript. F.K., G.N. and Z.D. assisted in the study design and examined the preliminary data. F.K., G.N. and Z.D. helped analyze the protease and bioinformatics output data. All authors assisted in the final assessment of data and reviewed the manuscript. Z.D. conceived the study design, obtained funding for the study, and revised and edited the manuscript. All authors have read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: El-Assaad, A. *et al.* Novel Bioinformatics–Based Approach for Proteomic Biomarkers Prediction of Calpain-2 & Caspase-3 Protease Fragmentation: Application to βII-Spectrin Protein. *Sci. Rep.* **7**, 41039; doi: 10.1038/srep41039 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.