SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPUTATIONAL BIOLOGY
BIOINFORMATICS
PROTEINS
POST-TRANSLATIONAL MODIFICATIONS

Received
20 December 2011

Accepted
23 January 2012

Published
16 February 2012

Correspondence and requests for materials should be addressed to S.T. (sun.tian@nuolan.net) or J.W. (wujianhua@scut.edu.cn)

* Visiting scientist at Institute of Biomechanics.

Present address: Agendia NV, Science Park 406, Amsterdam, The Netherlands.

Computational prediction of furin cleavage sites by a hybrid method and understanding mechanism underlying diseases

Sun Tian[1,2*#], Wang Huajun[2] & Jianhua Wu[1]

[1]Institute of Biomechanics, School of Bioscience and Bioengineering, South China University of Technology, Guangzhou Higher Education Mega Centre, Panyu Distric, Guangzhou 510640, China, [2]Nuolan Net, Amsterdam, The Netherlands.

Furin cleaves diverse types of protein precursors in the secretory pathway. The substrates for furin cleavage possess a specific 20-residue recognition sequence motif. In this report, based on the functional characterisation of the 20-residue sequence motif, we developed a furin cleavage site prediction tool, PiTou, using a hybrid method composed of a hidden Markov model and biological knowledge-based cumulative probability score functions. PiTou can accurately predict the presence and location of furin cleavage sites in protein sequences with high sensitivity (96.9%) and high specificity (97.3%). PiTou's prediction scores are biological meaningful and reflect binding strength and solvent accessibility of furin substrates. A prediction result is interpreted within cellular contexts: subcellular localisation, cellular function and interference by other dynamic protein modifications. Combining next-generation sequencing, PiTou can help with elucidating the molecular mechanism of furin cleavage-associated human diseases. PiTou has been made freely available at the associated website.

Furin cleaves inactive protein precursors in the secretory pathway and controls the activation of diverse types of functional proteins[1,2]. The known substrates that are activated by furin include both host proteins and pathogen proteins. The biological functional categories of furin substrates are diverse and include extracellular matrix proteins, signalling peptides, hormone, growth factors, serum proteins, transmembrane receptors, ion channels, bacterial toxins and viral fusion peptides[3]. Regulation of furin-mediated substrate cleavage plays a crucial role in embryogenesis, pathogen infection, neurologic disease and cancer[3]. In addition, the utility of furin cleavage-targeted selective anti-cancer drug delivery is also being explored[4].

The execution of furin cleavage depends on the recognition of the furin cleavage site motif by the functional furin enzyme. The furin cleavage site motif was initially described as a four amino acid pattern: R-X-[K/R]-R↓[1]. However, this pattern does not explain all furin cleavage sites, e.g. the furin cleavage sites of the human albumin precursor VFRR↓DA[5] and the human C-type natriuretic peptide precursor RLLR↓ DL[6] cannot be described by the pattern R-X-[K/R]-R↓. On the other hand, a mutated form of Sindbis Virus PE2 protein RSKR↓LV contains the pattern R-X-[K/R]-R↓ but is not efficiently cleaved by furin[7]. In our previous work, the furin cleavage site was re-analysed and characterised as a 20 amino acid motif running from position P14 to position P6′, which can be divided into one core region (eight amino acids from P6–P2′) and two flanking solvent accessible regions (eight amino acids from P7–P14 and four amino acids from P3′–P6′)[8]. The core region (P6–P2′) fits into to the furin catalytic pocket and determines the binding strength. The flexible solvent accessible regions (P7–P14 and P3′–P6′) flank the core region. They provide the accessibility of the core region to the furin binding pocket and also facilitate conformational changes of the core region required by the dynamic furin cleavage process.

Our previous analysis indicated that the physical properties of this 20-residue motif are evolutionarily conserved across different organisms, including mammals, bacteria and viruses[8,9]. Furthermore, the biology underlying the relationship between the physical properties of furin cleavage sites, cellular function and viral infectivity has been analysed[8]. FurinDB, a database of 20-residue furin cleavage sites and associated drugs, was then constructed to provide a solid publicly available infrastructure for furin cleavage-related studies[10]. The functionally characterised 20-residue motif of the furin cleavage recognition site and FurinDB laid down an important theoretical foundation for the development of a reliable prediction tool for furin cleavage sites. In this report, we

developed a furin cleavage site prediction tool: PiTou. PiTou can predict the presence and location of furin cleavage site on protein sequences. PiTou is designed based on the functional characterisation of the underlying biology of furin cleavage site motifs. The PiTou algorithm is implemented as a hybrid method that combines the advantages of both a machine learning-based hidden Markov model and a set of biological mechanism-based cumulative probability score functions. The performance of the prediction tool is high, with a sensitivity of 96.9% and specificity of 97.3%. PiTou's prediction scores are biological meaningful, and they reflect binding strength and solvent accessibility of furin substrates. A prediction result also need to be interpreted within biological meaningful cellular contexts: subcellular localisation, cellular function and interference by other dynamic protein modifications. Combining next-generation sequencing, PiTou can help to discover the molecular mechanism underlying furin cleavage site-associated human diseases. PiTou has been made freely available at the associated website http://www.nuolan.net/reference.html.

Results

Performance of PiTou on the prediction of furin cleavage sites, its designing features and comparison with the other prediction method. Cross-validation is proven to be an effective method of evaluating the predictive performance of a prediction tool[11]. Leave-one-out cross-validation (LOOCV) was used to evaluate the sensitivity of PiTou on 131 known furin cleavage sites. The sensitivity from the cross-validation reached 96.9% (F_n false negative rate 3.1%, 127 out of total 131 sites) (Supplemental Information S1). The specificity was estimated using 4265 arginine sites that are not cleaved by furin in the experiments. A specificity of 97.3% (F_p false positive rate 2.7%, 4151 out of total 4265 sites) (Supplemental Information S2) was reached. The detailed results of cross-validation and specificity estimation are available as supplementary materials on the web site.

The performance of PiTou was compared with the published furin cleavage prediction tool ProP method[12]. The sensitivity and specificity values of ProP were taken from the original paper. Compared with ProP, PiTou showed superior sensitivity and specificity (**Table 1**). Unfortunately, a direct comparison of sensitivity and specificity of PiTou method and ProP method on a same independent test dataset is currently not possible due to two limitations: (1) the cross validation implementations of tools are not publicly available; (2) the number of known furin cleavage sites is small, and no independent testing dataset of furin cleavage sites is available and can be used for comparing these two tools. Therefore, the sensitivity and specificity values taken from ProP publication and those of PiTou may not be easily compared. However, the designing features of PiTou is evidently distinct from that of ProP. The high sensitivity and high specificity of the PiTou furin cleavage site prediction tool benefitted from its designing features. The most important feature is that PiTou is entirely biological knowledge-based, but the prediction score is substantiated by a machine learning-based hidden Markov model and cumulative probability score functions (**Methods**). This designing feature evolved from understanding the 20-residue furin cleavage site motif responsible for recognition by furin[8]. The constraints imposed by the physical properties of the 20-residue furin cleavage motif and the 3D binding model of substrates to the furin catalytic domain were translated into the score functions of PiTou (**Figure 1**). Biological and structural information on this 20-residue motif enhance our understanding of molecular biology of furin cleavage and thus improve the PiTou's prediction accuracy. On the contrary, the ProP method's designing feature is very different. The ProP method is a pure machine learning based method that entirely relies on automatic training process of neural networks[12]. The neural networks of ProP method consider the biology of furin cleavage and binding of substrates to the furin binding pocket as a black box (**Table 1**).

Integrating next generation sequence analysis and elucidating the molecular mechanisms of furin cleavage site-associated human diseases. Next-generation sequencing can rapidly sequence hundreds of genes and identify genomic mutations in the onset of human diseases or mutations which have emerged in the progression of a human disease, such as cancer[13]. With the accumulation of large amount of genomic data, one big challenge is to understand the cellular functional consequence of genetic mutations identified. A change in an amino acid in the 20-residue motif of a furin cleavage site can alter the pattern of the favoured physical properties of that specific position or region and thus affect the furin cleavage efficiency on the substrate. Malfunctions in furin cleavage efficiency will in turn cause human disease. Three examples of mutations resulting a loss or gain of furin cleavage sites are known, and almost all the known examples are associated with the molecular mechanism of a human disease: X-linked hypohydrotic ectodermal dysplasia, arrhythmogenic right ventricular dysplasia/cardiomyopathy disorder, prolonged thrombin time and a mild bleeding tendency (**Table 2**).

In all three cases illustrated in table 2 (**Table 2**), PiTou successfully predicted the loss of known functional furin cleavage sites or gain of not naturally occurring furin cleavage sites as the consequence of genomic mutations, thus in turn identified the molecular mechanism underlying some furin cleavage associated human diseases. In addition, all three examples showed that mutations at positions around furin cleavage sites can have dramatic consequences and cause diseases and disorders. Particularly, these examples provide three interesting insights:

1) Disease can be caused not only by the loss of a normal functional furin cleavage site, but also the gain of an aberrant furin cleavage site.
2) Both loss and gain of furin cleavage do not necessarily require mutations at the arginine at position P1 or P4.
3) A mutation directly present in the catalytic domain or the regulatory domain on a protein is not the only way that a genetic mutation can cause cellular functional consequence. A genetic mutation can also alter a short sequence motifs required by the interaction with other proteins, and result in entirely different cellular phenotype. This concept was demonstrated by the genomic mutations resulted changes of the proteolytic cleavage of extracellular enzymes by furin.

Discussion

Given the high sensitivity (96.9%) and high specificity (97.3%), PiTou can be used to identify potential furin cleavage sites on various types of extracellular proteins, e.g. extracellular matrix proteins,

Table 1 | Comparison of performance and design feature of furin cleavage site prediction tools

Prediction tool	Sensitivity	Specificity	Design method	Reference
PiTou	96.9%	97.3%	Biological knowledge-based: combination of a hidden Markov model and biological knowledge-based cumulative probability score functions	Results section
ProP	94.7%	83.7%	Pure machine learning neural network	Duckert *et al.* 2004[12]

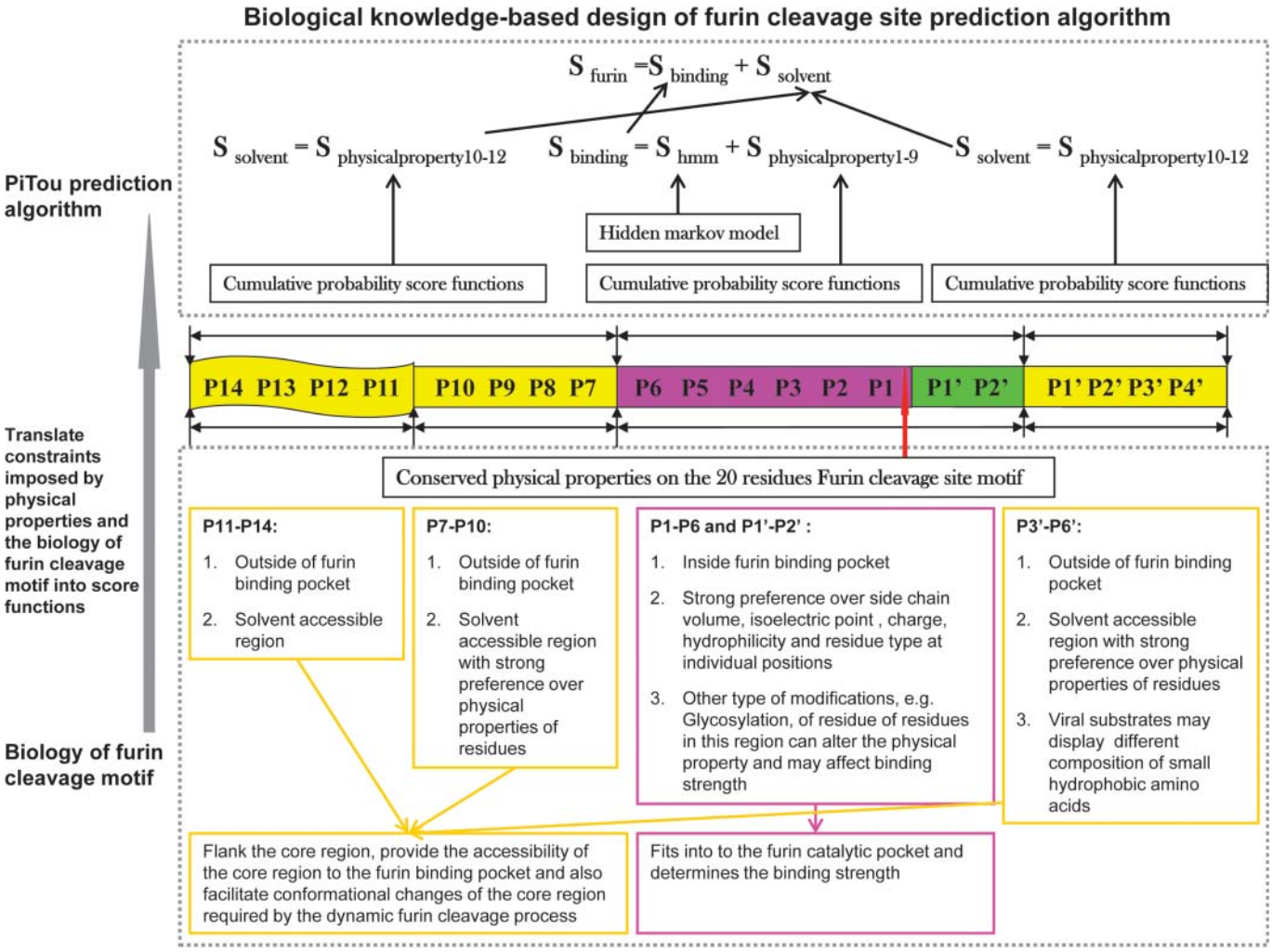

Figure 1 | The design of PiTou algorithm: the PiTou score function S_{furin} is biological knowledge-based and it comprises of a machine learning-based hidden Markov model and a set of biological mechanism-based cumulative probability score functions. S_{furin} is the sum of two parts: the core region binding score $S_{Binding}$ (calculated from eight amino acids at P6–P2′) and the flanking region solvent accessible score $S_{Solvent}$ (calculated from eight amino acids at P7–P14 and four amino acids at P3′–P6′). $S_{Binding}$ and $S_{Solvent}$ incorporate the physical properties of the 20-residue furin cleavage motif and the binding of substrates to the furin catalytic domain. The analysis of the 20-residue furin cleavage motif is described in previous publication[8].

signalling peptides, hormones, growth factors, serum proteins, transmembrane receptors, ion channels, bacterial toxins and viral proteins. The predicted cleavage sites will provide experimental groups with directions for the elucidation of the possible underlying cellular mechanisms of the observed cellular phenotype.

The final prediction score S_{furin} of PiTou is an implementation of two main criteria (**Figure 1**):

a) $S_{Binding}$: the binding strength of the core region to the furin catalytic pocket
b) $S_{Solvent}$: the accessibility of the core region to the furin catalytic pocket, supported by the hydrophilicity and flexibility of the flanking regions.

The biological knowledge about the binding complex is reflected by $S_{PhysicalProperty}$ which contributes to both $S_{Binding}$ and $S_{Solvent}$. As a consequence, by examining the output of the core region binding score $S_{Binding}$ and the flanking region solvent accessible score $S_{Solvent}$, PiTou not only predicts whether an arginine site is cleaved by furin, but also predicts the molecular mechanism by which furin cleavage can or cannot take place at a given arginine site, e.g. because the binding strength of core region is weak, resulting in low $S_{Binding}$ or because the accessibility of the core region is poor, resulting in low $S_{Solvent}$, etc. This design feature allows the possibility of using the PiTou score set S_{furin} as an indicator to engineer the amino acids around a given arginine site with the 20-residue motif pattern[8] to increase/decrease the binding strength score $S_{Binding}$ or the accessibility score $S_{Solvent}$, thus increasing/decreasing desired furin cleavage efficiency, or to an extreme extent, to create/diminish a furin cleavage site that can possibly modify cell biology under disease conditions. This is conceptually different from pure machine learning-based ProP approaches that produce a prediction result, while our knowledge of the molecular mechanism of furin cleavage process is considered to be a black box, thus does not necessarily improved.

The positive prediction result of PiTou $S_{furin} > 0$ indicates that an arginine site can be cleaved by furin; however, a prediction result needs to be interpreted within biological meaningful cellular contexts. There are three cellular contexts which still need to be considered:

1. The accessibility of furin cleavage sites in the context of subcellular localisation. Functional furin does not come into contact with cytosolic proteins because furin is an extracellular enzyme. Equally, for potential furin cleavage recognition sites present on transmembrane proteins, the cytosolic part of a transmembrane protein should not have the opportunity to come into contact with functional furin, whereas only the extracellular part of

Table 2 | Predictions of PiTou on furin cleavage sites to elucidate the molecular mechanisms of furin cleavage-associated human diseases.

Protein	Furin cleavage site motif	PiTou prediction score	Genetic mutation	Resulting amino acid mutation	Position of resulting amino acid mutation on furin cleavage site motif	Mutated furin cleavage site motif	PiTou prediction score of mutated furin cleavage site motif	Cellular consequence	Associated Human Disease
Ectodysplasin-A isoform II	EKPYSEEESRRVRR↓NKRSKS	15.9 (cleaved by furin)	C704->T	R155->C	P2: R->C	EKPYSEEESRRVCR↓ NKRSKS	−9.8	Loss of furin cleavage site	X-linked hypohydrotic ectodermal dysplasia[21,22]
			C707->T	R156->C	P1: R->C	EKPYSEEESRRVCC↓ NKRSKS	Loss of P1 arginine	Loss of furin cleavage site	
			G708->A	R156->H	P1: R->H	EKPYSEEESRRVCH↓ NKRSKS	Loss of P1 arginine	Loss of furin cleavage site	
	YSEEESRRVRRNKR↓SKSNEG	17.0 (cleaved by furin)	C704->T	R155->C	P5: R->C	YSEEESRRVCRNKR↓ SKSNEG	−8.3	Loss of furin cleavage site	
			C707->T	R156->C	P4: R->C	YSEEESRRVRCNKR↓SKSNEG	−10.3	Loss of furin cleavage site	
			G708->A	R156->H	P4: R->H	YSEEESRRVRHNKR↓SKSNEG	9.3	Reduced furin cleavage efficiency	
Desmoglein 2	KLLPKHPHLVRQKR↓ AWITAP	10.6 (cleaved by furin)	G143->A	R48->H	P1: R->H	KLLPKHPHLVRQKH↓ AWITAP	Loss of P1 arginine	Loss of furin cleavage site	Arrhythmogenic right ventricular dysplasia/ cardiomyopathy disorder[23]
			G134->A	R45->Q	P4: R->Q	KLLPKHPHLVQQKR↓ AWITAP	−14.6	Loss of furin cleavage site	
Fibrinogen alpha chain precursor	GDFLAEGGGVRGPR↓ VVERHQ	−14.6 (not cleaved by furin)	T1215->A	V39->D	P1′: V->D	GDFLAEGGGVRGPR↓ DVERHQ	8.1	Gain of unwanted furin cleavage site	The mutated fibrinogen alpha chain was cleaved by furin. Prolonged thrombin time and a mild bleeding tendency[24].

transmembrane proteins can be cleaved by furin in the secretory pathway. Subcellular localisation of a protein can be used to eliminate of false positive predictions.

2. Cellular functions of substrates. The most interesting examples are viral fusion peptides. The P3′–P6′ region of viral fusion peptides cannot be too hydrophilic or too hydrophobic because the virus appears to need a subtle balance between sufficient hydrophobicity required by the viral fusion process and sufficient hydrophilicity required by furin cleavage efficiency[8]. This biology of viral fusion with furin cleavage present an contradicted logic and unique challenge for the life cycle of virus. The motif analysis suggested that viruses have cunningly solved this dilemma by tuning the composition of small hydrophobic amino acids (glycine, alanine and proline) in the P3′–P6′ region[8]. As a consequence of the presence of hydrophobic amino acids, the average hydrophobicity scale of the P3′–P6′ region of these viral substrates is much higher than the average of mammalian and bacterial substrates (**Figure 2**, 0.005 versus −1.235, student t-test pvalue = 1.3E-004, calculated using the physical property EISD840101 consensus normalized hydrophobicity scale for amino acids)[8,14]. This hydrophobic stretch in the P3′–P6′ region of furin cleavage sites on viral spike proteins can sometimes result in a lower predicted score of PiTou. Therefore, a more careful consideration should be given when a query sequence has a viral origin.
3. The interference of furin cleavage by other dynamic modifications on an amino acid. This is a novel and interesting issue. An algorithm takes one letter symbol to represent an amino acid, e.g. A for alanine, R for arginine, S for serine, etc. However, the physical property of the same amino acid can be different under different conditions, and the very same type of amino acid may lead to completely different cellular consequences. The statistical analysis of the physical properties of a substrate in the P1′ position indicated that the total volume of a substrate fitting into position P1′ could not exceed the total volume available in the narrow furin binding pocket at position P1′, and therefore position P1′ has preference for small hydrophilic residues such as serine[8]. Three small residues, i.e. serine, threonine and asparagine, have the potential to be glycosylated. The volume of serine, threonine and asparagine will increase after N-linked glycosylation on asparagine and O-linked glycosylation on serine or threonine. The glycosylated amino acids no longer possess the preferred physical properties required at the P1′position. The same type of amino with the symbol N (asparagine) present at position P1′ can result in either efficient furin cleavage or defective furin cleavage, which can entirely depend on whether the side chain volume of N (asparagine) has been modified or not[7]. Therefore, the presence of small residues such as asparagine, serine or threonine at substrate position P1′ deserves particular attention in the analysis of furin cleavage-mediated viral infection. A sequence motif may be cleaved by furin in one cellular context (no glycosylation at position P1′), but the very same sequence motif may not be cleaved by furin in a different cellular context (glycosylation at position P1′).

The three discussions above (subcellular localisation of the region of substrates with transmembrane regions, hydrophobic tendencies in the P3′–P6′ region of viral fusion peptides and glycosylation interference with furin cleavage) showed that the cellular context and biological background are important for the interpretation of prediction results from a bioinformatics tool. Furthermore, they emphasise the important concept of studying the underlying molecular mechanism accompanying the design of a computational prediction tool rather than purely relying on analysing the sequence motif pattern with statistical models or machine learning-based

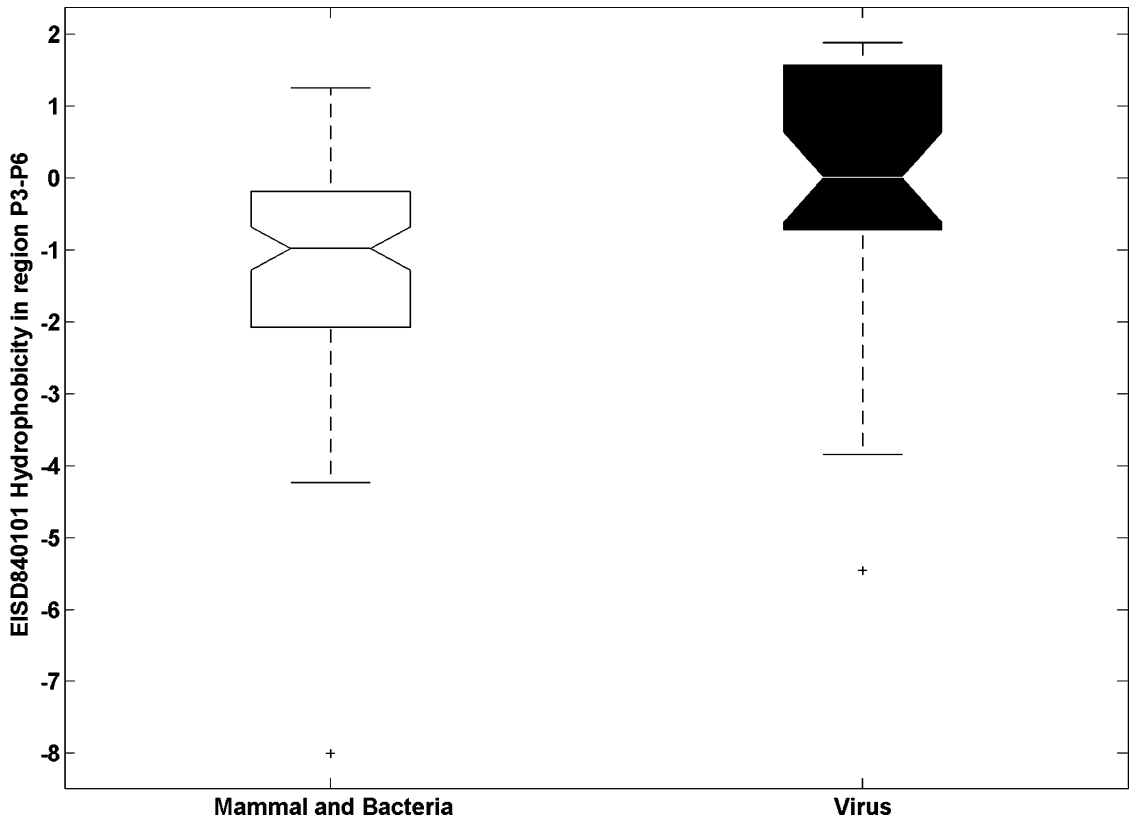

Figure 2 | The hydrophobicity scale of the P3′–P6′ region of viral substrates (filled black box) is much higher than that of mammalian and bacterial substrates (white box), student t-test pvalue = 1.3E-004.The hydrophobicity is calculated using the physical property EISD840101 consensus normalized hydrophobicity scale for amino acids[14].

methods. This concept not only limits to the development of furin cleavage site prediction tool, but may also apply to the development of other types of prediction tools such as clinical utility-related molecular diagnostic signatures[15].

PiTou has been demonstrated helpful for elucidating cellular functional consequence of genomic mutations and understanding the molecular mechanisms of furin cleavage site-associated human diseases. The number of genomic mutations resulting a loss or gain of a furin cleavage site associated with human disease may be underestimated. Because the furin cleavage site motif is comprised of about 20 residues (P14-P6′), not four residues (P4-P1) as previously thought, in theory, any mutation, deletion or insertion within the residues in these 20 positions can change the physical properties; this raises the possibility of losing or gaining a furin cleavage site or at least affecting furin cleavage efficiency. The PiTou furin cleavage prediction tool can serve as an efficient computational tool to screen and evaluate the possibility of an aberrant gain or loss of furin cleavage in the mutated protein sequence in patients with various disorders or diseases and thus help with elucidating the molecular mechanism of human diseases. Next-generation sequencing can identify thousands of genomic mutations in the progression of human diseases. PiTou can predict the functional consequence of these genomic mutations on furin cleavage efficiency. By combining next-generation sequencing and the PiTou furin cleavage site prediction tool, our fundamental understanding of the pathogenesis of human diseases has been enhanced.

The PiTou package is publicly available for download at www.nuolan.net/reference.html. We believe PiTou will provide a valuable publicly available computational tool to scientists in the field of molecular biology and molecular medicine.

Methods

Dataset of known furin cleavage sites. A dataset of 131 known furin cleavage sites was retrieved from FurinDB, a database of 20-residue furin cleavage sites, substrates and associated drugs[10]. Each site included is supported by experimental biochemical evidence[10]. The taxa of these substrates cover three different origins: virus, bacterial and mammals. All substrates included are cleaved by mammalian furin. The cellular function of the substrates in the dataset covers a representative functional spectrum: extracellular matrix proteins, signalling peptides, hormones, growth factors, serum proteins, transmembrane receptors, ion channels, bacterial toxins and viral fusion peptides[3]. The remaining 4265 arginine sites presented in the protein sequences of furin substrates but not reported to be cleaved by furin in the experiments were also collected as negative sites.

The set of 20 residues in the furin cleavage site were formatted and aligned into a multiple sequence alignment ($n_{furinSites} = 131$). One important consideration is the cell biology of furin cleavage. Furin is an extracellular enzyme and furin cleavage takes place after the secreted signal peptide of a protein sequence is cleaved off. For a host protein precursor, when the location of a furin cleavage site is very close to the N-terminal, the overlapping region between the known secreted signal peptides and the P1–P14 position of the furin cleavage site motif is substituted with gap symbols in the multiple sequence alignment.

Constructing the biological knowledge-based score function. PiTou is a biological knowledge-based furin cleavage site prediction tool that employs both a machine learning-based hidden Markov model and a set of biological knowledge-based cumulative probability score functions. The PiTou score function S_{furin} is the sum of two parts, similar as the scheme for the knowledge-based prediction of short functional motifs[16]: the core region binding score $S_{Binding}$ (calculated from eight amino acids at P6–P2′) and the flanking region solvent accessible score $S_{Solvent}$ (calculated from eight amino acids at P7–P14 and four amino acids at P3′–P6′). The overview of the design of PiTou algorithm is illustrated in figure 1 (**Figure 1**).

$$S_{Furin} = S_{Binding} + S_{Solvent}$$

$$S_{Binding} = S_{hmm} + \sum_{i=1}^{n} f_i * S_{PhysicalProperty}(i)$$

$$S_{Solvent} = \sum_{i=1}^{n} f_i * S_{PhysicalProperty}(i)$$

Hidden markov model (HMM) provide robust a probabilistic model that comprise of states with emission probabilities and transition probabilities. A HMM can sensitively measure the similarity between residues in a query protein sequence with homologous residues in a target set of protein sequence[17]. A profile hidden Markov model $FurinProfile_{hmm}$ is constructed using the multiple sequence alignment of furin cleavage sites. $FurinProfile_{hmm}$ evaluates the similarity of the core region (P6–P2′) of a query sequence to the amino acid type occurrence frequency in the core regions (P6–P2′) of known furin cleavage site sequences. The score S_{hmm} is the standard log-odd probabilities from this hidden Markov model $FurinProfile_{hmm}$[18].

$S_{PhysicalProperty}$ is a physical property score, each $S_{PhysicalProperty}$ score is calculated from a known physical property feature or biological feature presented on the furin cleavage site 20-residue motif[8]. Physical property values are retrieved from an AAindex database that stores various physical and biochemical properties of amino acids[19]. The 20-residue furin cleavage recognition site motif was formularised into 12 $S_{PhysicalProperty}$ functions (**Table 3**). Each $S_{PhysicalProperty}$ results in a negative score or a zero score. The absolute value of a negative score reflects the degree of deviation from the evolutionally conserved physical property pattern in the 20-residue furin cleavage site motif: a larger deviation results in a large negative value and a smaller deviation results in a small negative value. There are two types of $S_{PhysicalProperty}$ functions[16, 20]:

(1) Fixed value function type: $S_{PhysicalProperty}(i)$ is assigned as 1 if $PhysicalProperty_{p_1-p_n}(i)$ exceeds or is below a predefined threshold; otherwise, $S_{PhysicalProperty}(i)$ is assigned as 0. The predefined threshold of a fixed penalty $S_{PhysicalProperty}(i)$ is calculated from the known furin cleavage sites.

(2) Normal cumulative distribution function type: $S_{PhysicalProperty}(i)$ is the *log* probability from a normal cumulative distribution function. *Equation* $S_{PhysicalProperty}$ *1* was used for calculating $S_{PhysicalProperty}$ for the isoelectric point, charge and flexibility; *Equation* $S_{PhysicalProperty}$ *2* was used for calculating $S_{PhysicalProperty}$ for hydrophobicity and volume.

$$\Delta = PhysicalProperty_{p_1-p_n}(i) - \bar{X}_{KnownFurinSite}\left(PhysicalProperty_{p_1-p_n}\right)$$

Equation $S_{PhysicalProperty}$ *1 :*

Table 3 | List of 12 $S_{PhysicalProperty}$ functions that evaluate $S_{Binding}$ binding strength of the core region (P6–P2′) and $S_{Solvent}$ solvent accessibility of two flanking regions (P7–P14 and P3′–P6′)

$S_{PhysicalProperty}$ functions	Physical property	Position on the furin cleavage site motif	Description[8]
$S_{PhysicalProperty}$ 1	ZIMJ680104[25]	P2 P4 P5 P6	Positive charge and isoelectric point
$S_{PhysicalProperty}$ 2	ZIMJ680104[25]	P2 P3	Positive charge and isoelectric point
$S_{PhysicalProperty}$ 3	Cysteine[8]	P2–P6	Disulfide bond formation potential and negative charge
$S_{PhysicalProperty}$ 4	ZIMJ680104[25], EISD840101[14], KUHL950101[26]	P4	Aliphatic residue or positively charged residue
$S_{PhysicalProperty}$ 5	FAUJ880111[27]	P4–P6	Positive charge compensation
$S_{PhysicalProperty}$ 6	BULH740102[28]	P1′	Volume
$S_{PhysicalProperty}$ 7	BULH740102[28]	P1′ – P3′	Volume
$S_{PhysicalProperty}$ 8	KARP850103[29]	P1′ P2 P4 P5 P6	Flexibility
$S_{PhysicalProperty}$ 9	KARP850103[29]	P1′ – P3′	Flexibility
$S_{PhysicalProperty}$ 10	EISD840101[14]	P7 – P10	Hydrophobicity
$S_{PhysicalProperty}$ 11	EISD840101[14]	P3′–P6′	Hydrophobicity
$S_{PhysicalProperty}$ 12	EISD840101[14]	P11–P14	Hydrophobicity

$$S_{PhysicalProperty}(i) = \left\{ log\left(normcdf\left(\frac{PhysicalProperty_{p_1-p_n}(i) - \overset{0,\Delta\geq 0}{\bar{X}_{KnownFurinSite}}\left(PhysicalProperty_{p_1-p_n}\right)}{\delta_{KnownFurinSite}\left(PhysicalProperty_{p_1-p_n}\right)} \right) \right), \Delta < 0 \right.$$

Equation $S_{PhysicalProperty}$ *2 :*

$$S_{PhysicalProperty}(i) = \left\{ log\left(normcdf\left(\frac{\overset{0,\Delta\leq 0}{\bar{X}_{KnownFurinSite}}\left(PhysicalProperty_{p_1-p_n}\right) - PhysicalProperty_{p_1-p_n}(i)}{\delta_{KnownFurinSite}\left(PhysicalProperty_{p_1-p_n}\right)} \right) \right), \Delta > 0 \right.$$

$S_{PhysicalProperty\ 1-9}$ evaluate the potential binding strength of the core region to the furin catalytic pocket and contribute to the binding score $S_{Binding}$. $S_{PhysicalProperty\ 10-12}$ evaluate the potential accessibility of the core region to the furin catalytic pocket and contribute to the flanking region solvent accessible score $S_{Solvent}$.

All scores S_{Furin} , $S_{Solvent}$, $S_{Binding}$, S_{hmm} , $S_{PhysicalProperty}$ are log-odd probabilities. Every single arginine site presented in the query protein sequence is considered as a potential furin cleavage site and the 20-residue sequence motif encompassing this arginine was evaluated using the prediction score function S_{furin}. An arginine site with a predicted score $S_{furin} \geq 0$ is interpreted as a predicted furin cleavage site.

1. Nakayama, K. Furin: a mammalian subtilisin/Kex2p-like endoprotease involved in processing of a wide variety of precursor proteins. *Biochem. J.* **327 (Pt3),** 625–635 (1997).
2. Molloy, S. S., Anderson, E. D., Jean, F. & Thomas, G. Bi-cycling the furin pathway: from TGN localization to pathogen activation and embryogenesis. *Trends Cell Biol.* **9**, 28–35 (1999).
3. Thomas, G. Furin at the cutting edge: from protein traffic to embryogenesis and disease. *Nat. Rev. Mol. Cell Biol.* **3**, 753–766 (2002).
4. Hajdin, K., D'Alessandro, V., Niggli, F. K., Schafer, B. W. & Bernasconi,M. Furin targeted drug delivery for treatment of rhabdomyosarcoma in a mouse model. *PLoS. One.* **5**, e10445 (2010).
5. Brennan, S. O. & Nakayama, K. Furin has the proalbumin substrate specificity and serpin inhibitory properties of an in situ hepatic convertase. *FEBS Lett.* **338**, 147–151 (1994).
6. Wu, C., Wu, F., Pan, J., Morser, J. & Wu, Q. Furin-mediated processing of Pro-C-type natriuretic peptide. *J. Biol. Chem.* **278**, 25847–25852 (2003).
7. Klimstra, W. B., Heidner, H. W. & Johnston, R. E. The furin protease cleavage recognition sequence of Sindbis virus PE2 can mediate virion attachment to cell surface heparan sulfate. *J. Virol.* **73**, 6299–6306 (1999).
8. Sun, T. A 20 Residues Motif Delineates the Furin Cleavage Site and its Physical Properties May Influence Viral Fusion. *Biochemistry Insights* **2**, 9–20 (2009).
9. Sun, T. & Wu, J. Comparative study of the binding pockets of mammalian proprotein convertases and its implications for the design of specific small molecule inhibitors. *Int J Biol Sci* **6**, 89–95 (2010).
10. Sun, T., Huang, Q., Fang, Y. & Wu, J. FurinDB: A Database of 20-Residue Furin Cleavage Site Motifs, Substrates and Their Associated Drugs. *Int. J. Mol. Sci.* **12**, 1060–1065 (2011).
11. Cawley, G. C. & Talbot, N. L. Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. *Neural Netw.* **17**, 1467–1475 (2004).
12. Duckert, P., Brunak, S. & Blom,N. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel* **17**, 107–112 (2004).
13. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* **55**, 641–658 (2009).
14. Eisenberg, D. Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* **53**, 595–623 (1984).
15. Sun, T. *et al.* Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. *Biomark. Insights.* **5**, 129–138 (2010).
16. Neuberger, G. A Framework for the Knowledge-Based Prediction of Short FunctionalMotifs from Amino Acid Sequence. 2006. Vienna university. Ref Type: Thesis/Dissertation.
17. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
18. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge University Press,1998).
19. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **27**, 368–369 (1999).
20. Sun, T. Sequence-analytic characterization and prediction of furin cleavage recognition site based on a simple substrate-catalytic domain binding structural model. 2007. Graz University of Technology. Ref Type: Thesis/Dissertation.
21. Vincent, M. C., Biancalana, V., Ginisty, D., Mandel, J. L. & Calvas, P. Mutational spectrum of the ED1 gene in X-linked hypohidrotic ectodermal dysplasia. *Eur. J. Hum. Genet.* **9**, 355–363 (2001).
22. Chen, Y. *et al.* Mutations within a furin consensus sequence block proteolytic release of ectodysplasin-A and cause X-linked hypohidrotic ectodermal dysplasia. *Proc. Natl. Acad. Sci. U. S. A* **98**, 7218-7223 (2001).
23. Awad, M. M. *et al.* DSG2 mutations contribute to arrhythmogenic right ventricular dysplasia/cardiomyopathy. *Am. J. Hum. Genet.* **79**, 136–142 (2006).
24. Brennan, S. O., Hammonds, B. & George, P. M. Aberrant hepatic processing causes removal of activation peptide and primary polymerisation site from fibrinogen Canterbury (A alpha 20 Val --> Asp). *J. Clin. Invest* **96**, 2854–2858 (1995).
25. Zimmerman, J. M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **21**, 170–201 (1968).
26. Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A. & Getzoff, E. D. Atomic and residue hydrophilicity in the context of folded protein structures. *Proteins* **23**, 536–547 (1995).
27. Fauchere, J. L., Charton, M., Kier, L. B., Verloop, A. & Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278 (1988).
28. Bull, H. B. & Breese, K. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* **161**, 665–670 (1974).
29. Karplus, P. A. & Schulz, G. E. Prediction of Chain Flexibility in Proteins - A Tool for the Selection of Peptide Antigens. *Naturwissenschaften* **72**, 212–213 (1985).

Acknowledgement

This work was stared as Sun Tian's PhD thesis at TUGraz and fund by GENAU Bioinformatics Integration Network PhD programme (2005–2007) and NSFC grant 11072080. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Wang Huajun and Sun Tian thank Mr.Deng Huixian for his inspiration.

Author contributions

Sun Tian and Jianhua Wu designed and implemented C++ version of the algorithm, finalised and compared different machine learning methods. Sun Tian and Wang Huajun implemented algorithm, wrote the PiTou software package codes and tested it. Sun Tian provided additional analysis of biology related to PiTou prediction. Sun Tian draft the manuscript. All authors reviewed the manuscript.

Additional information

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

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

How to cite this article: Tian, S., Huajun, W. & Wu, J. Computational prediction of furin cleavage sites by a hybrid method and understanding mechanism underlying diseases. *Sci. Rep.* 2, 261; DOI:10.1038/srep00261 (2012).