SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MACHINE LEARNING
COMPUTATIONAL MODELS
TARGETED THERAPIES
TUMOUR ANGIOGENESIS

Received
23 November 2012

Accepted
22 March 2013

Published
5 April 2013

Correspondence and requests for materials should be addressed to G.P.S.R. (raghava@imtech.res.in)

* These authors contributed equally to this work.

Computational approach for designing tumor homing peptides

Arun Sharma*, Pallavi Kapoor*, Ankur Gautam, Kumardeep Chaudhary, Rahul Kumar, Jagat Singh Chauhan, Atul Tyagi & Gajendra P. S. Raghava

Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh-160036, India.

Tumor homing peptides are small peptides that home specifically to tumor and tumor associated microenvironment *i.e.* tumor vasculature, after systemic delivery. Keeping in mind the huge therapeutic importance of these peptides, we have made an attempt to analyze and predict tumor homing peptides. It was observed that certain types of residues are preferred in tumor homing peptides. Therefore, we developed support vector machine based models for predicting tumor homing peptides using amino acid composition and binary profiles of peptides. Amino acid composition, dipeptide composition and binary profile-based models achieved a maximum accuracy of 86.56%, 82.03%, and 84.19% respectively. These methods have been implemented in a user-friendly web server, TumorHPD. We anticipate that this method will be helpful to design novel tumor homing peptides. TumorHPD web server is freely accessible at http://crdd.osdd.net/raghava/tumorhpd/.

Cancer is a major public health concern and remains a leading cause of mortality across the globe. This devastating disease affects both developed and developing countries. Despite the considerable progress in understanding the molecular basis of cancer, mortality rate is still high[1]. The chemotherapy is the principal mode of current cancer treatment, but it is limited by significant toxicity and frequently acquired resistance[2]. In the last decade, treatment options for cancer have shifted towards the more specific targeted therapies[3,4]. Many strategies have been exploited to target tumors. The most commonly used strategy is engineered antibodies or antibody fragments[5]. Though monoclonal antibodies are very selective, poor penetration inside the tumors and high production cost hinders their usage as therapeutic agents[6]. Nowadays, use of peptides for tumor targeting is getting much attention. In this context, tumor homing peptides (THPs) have become a very promising strategy to deliver therapeutics at tumor site. In the last decade, much attention has been paid on targeting tumor cells or tumor vasculature using THPs[7].

THPs are short peptides (3–15 amino acids), which specifically recognize and bind to tumor cells or tumor vasculature. Since the introduction of tumor homing concept in 1998, a large number of THPs have been identified by *in vitro* and *in vivo* phage display technology. THPs have some common motifs like RGD, NGR, which specifically bind to a surface molecule on tumor cells or tumor vasculature. For example, RGD peptide binds to α integrins[8] and NGR binds to a receptor aminopeptidase N, which is present on the surface of tumor endothelial cells[9]. Due to their tumor homing capability, THPs are being used in cancer diagnosis and treatment. Many anti-cancer drugs and imaging agents have been targeted to tumor site in mice models once conjugated with THPs[10]. The results of such studies are very encouraging and few THPs are already in clinical trials[11].

With such potential of THPs in cancer therapeutics, the computer aided prediction of THPs would be very beneficial in designing and developing novel THPs, thus saving time and labor of experimental biologists. To the best of authors' knowledge, no method has been developed for predicting/designing THPs. In the present study, a systematic attempt has been made to develop highly accurate support vector machine (SVM)-based models using various features of proteins/peptides like amino acid composition (AAC), dipeptide composition (DPC) and binary profile patterns (BPP). A user-friendly web server has also been developed to help the cancer biologists to predict and design THPs.

Results

Analysis of THPs. *Compositional analysis.* In order to find out overall dominant residues in THPs, we computed and compared percent amino acid composition of THPs and non-THPs in the main dataset. It was observed that certain types of residues like C, R, G, W, P, L and S are more abundant in THPs (Figure 1). In order to understand preference of residues at N- and C-terminals, we computed and compared percent AAC of N- and C-terminus

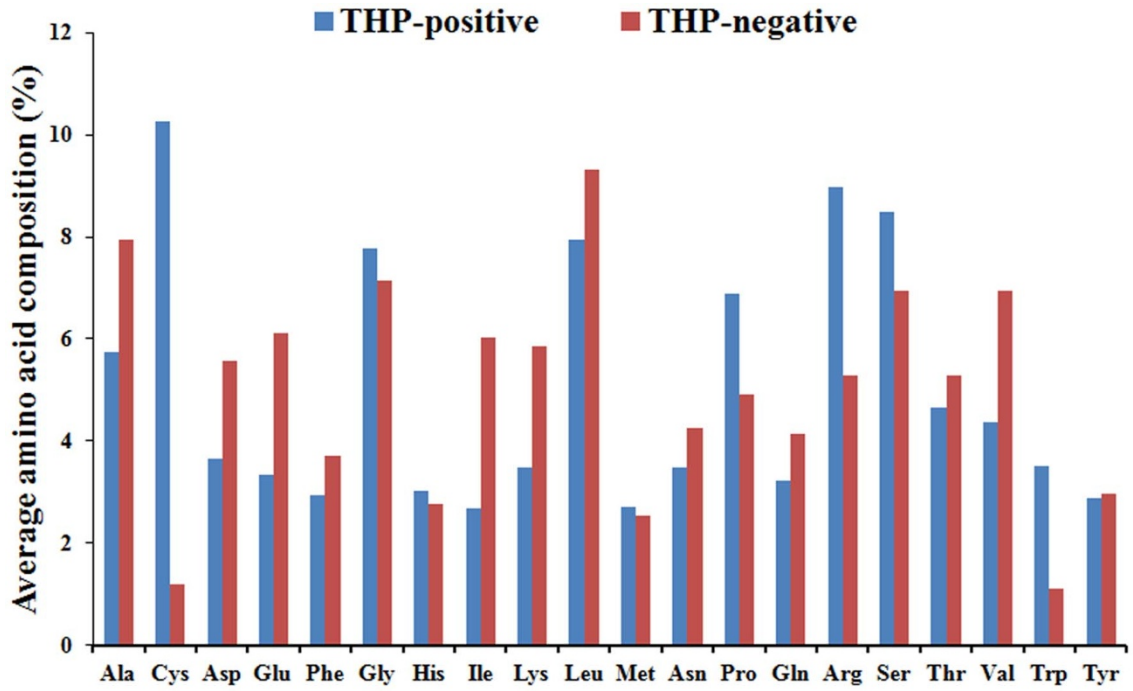

Figure 1 | Average amino acid composition. Comparison of percent average amino acid composition of peptides (THPs and non-THPs (randomly generated peptides)).

residues of THPs and non-THPs. However, we did not find any significant difference in AAC in terminal residues (data not shown).

Preference of residues. In order to understand preference of certain types of residues at different positions in THPs, we generated sequence logos. The sequence logos of 10 N-terminal and C-terminal residues of peptides are shown in Figure 2 & 3, respectively. As shown in Figure 2, certain residues are preferred at specific positions, *e.g.*, C, A, S, G at first position; G, R, P, E at 2nd position *etc.* Overall, THPs are dominated by certain type of residues like C, G, L, P *etc.*, being present at most of the positions. Similarly, certain residues are preferred at the C-terminus (Figure 3), for example, residues P, R, C, N and S are preferred at most of the positions.

AAC-based model. In compositional analysis of THPs, it has been observed that certain residues are dominated over others. This means that THPs and non-THPs can be discriminated on the basis of their AAC. Based on this observation, we developed SVM model on main dataset. The performance of AAC-based SVM model has been shown in Table 1. The model developed on main dataset achieved a maximum accuracy of 82.52% with an MCC and area under the curve (AUC) of 0.65 and 0.90 respectively. Similarly, SVM models were developed on subsets NT5, CT5, NTCT5, NT10, CT10, and NTCT10 and performances of these models have been summarized in Table 1. Model developed with NTCT10 dataset achieved a little higher accuracy of 86.56% with MCC 0.70 and AUC 0.91.

DPC-based model. DPC has been used previously to discriminate different classes of proteins[12]. Dipeptide encapsulates the global information of the amino acid fraction as well as the local order of amino acids. Thus, DPC is a better feature as compared to AAC. Therefore, SVM models based on DPC have been constructed on all

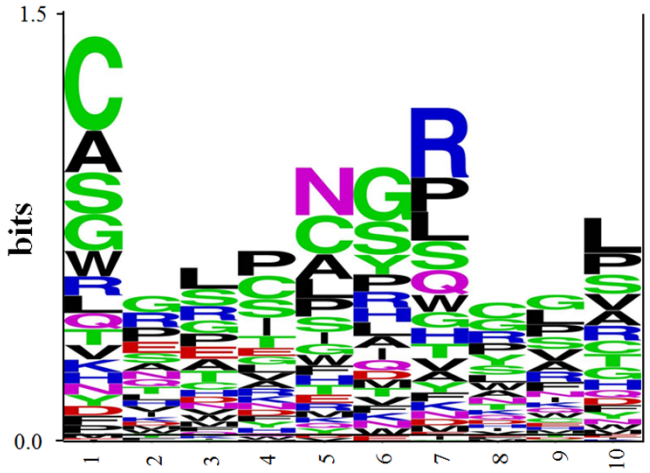

Figure 2 | Sequence logo of first ten residues (N-terminus) of THPs. The figure depicts the sequence logo of first ten residues (N-terminus) of THPs, where size of residue is proportional to its propensity.

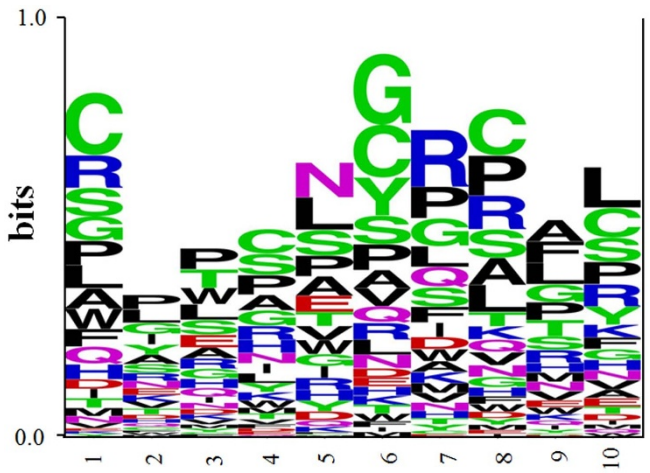

Figure 3 | Sequence logo of last ten residues (C-terminus) of THPs. The figure depicts the sequence logo of last ten residues (C-terminus) of THPs, where size of residue is proportional to its propensity.

Table 1 | Performances of SVM models developed using amino acid composition of peptides

Dataset	Sensivity	Specificity	Accuracy	MCC	AUC
Main dataset	81.57	83.46	82.52	0.65	0.90
NT5	70.88	83.41	77.15	0.55	0.83
CT5	70.57	81.72	76.15	0.53	0.82
NTCT5	77.2	84.95	81.08	0.62	0.88
NT10	78.66	79.45	79.05	0.58	0.86
CT10	84.19	78.26	81.23	0.63	0.87
NTCT10	80.63	89.71	86.56	0.70	0.91

*MCC: Matthew's correlation coefficient; AUC: Area under curve.

the datasets. Performances of DPC-based models are summarized in Table 2. Overall, performance of DPC-based models is poorer than AAC-based model. DPC-based model developed with main dataset achieved maximum accuracy of 81.29% with MCC and AUC values of 0.63 and 0.90 respectively, which is less than the models based on AAC (Table 2). Model developed with NTCT10 dataset achieved a maximum accuracy of 82.03% with MCC and AUC values of 0.63 and 0.88, respectively.

BPP-based method. In THPs, certain residues are preferred at specific positions on N- and C-terminus (Figure 2 & 3). Therefore, to implement the information about frequency as well as the order of residues, we made an attempt to develop a method using binary profiles of peptides. We have generated BPP of peptides. In binary pattern, a vector of dimension 20 represents a residue, and for *N* residues the input vector of dimension is 20 × N. We have used the following three approaches:

N-terminal approach: In this approach, we used subsets NT5 and NT10, consist of 5 and 10 N-terminal residues of THPs and non-THPs (See Material and Methods). We extracted 5 and 10 N-terminus residues from each peptide, and generated binary profile of dimension 5 × 20, and 10 × 20, respectively. These profiles were then used to develop SVM models. The accuracy of models developed on NT5 and NT10 datasets were 77.08% and 81.03% with MCC 0.54, 0.62 and AUC 0.84, 0.89 respectively (Table 3).

C-terminal approach: We adopted same strategy for the C-terminal as used for the N-terminal except taking the residues from C-terminal instead of N, using subsets CT5 and CT10. The performance of BPP-based SVM model using 5 and 10 C-terminal residues was almost similar to N-terminal approach. As shown in table 3, we achieved maximum accuracy of 76.38% and 79.84% with MCC of 0.53 and 0.60 for 5 and 10 C- terminal residues of peptides respectively.

N- and C-terminal approach: In order to check, if using the N- and C-terminal of the peptides together would enhance the accuracy of the method, we developed models using N- and C-terminal residues. In this approach, we made two subsets named NTCT5 and NTCT10. First model was developed using BPP of first 5 residues from N-terminal and 5 residues from C-terminal. Second model was developed using BPP of 10 residues from N-terminal and 10 residues from C-terminal. As shown in Table 3, we achieved maximum accuracy 84.19% with MCC 0.69 and AUC 0.91 for NTCT10 subset.

Table 2 | The performances of SVM models developed using dipeptide composition of peptides

Dataset	Sensitivity	Specificity	Accuracy	MCC*	AUC
Main dataset	83.87	78.71	81.29	0.63	0.90
NT5	71.49	72.81	72.15	0.44	0.79
CT5	71.49	72.35	71.92	0.44	0.79
NTCT5	63.48	89.25	76.38	0.55	0.85
NT10	79.84	75.1	77.47	0.55	0.85
CT10	74.7	81.03	77.87	0.56	0.84
NTCT10	84.58	80.67	82.03	0.63	0.88

*MCC: Matthew's correlation coefficient; AUC: Area under curve.

Table 3 | Performances of SVM models developed using binary profile of peptides

DATASET	Sensitivity	Specificity	Accuracy	MCC*	AUC
NT5	76.12	78.03	77.08	0.54	0.84
CT5	70.57	82.18	76.38	0.53	0.83
NTCT5	74.88	87.25	81.08	0.63	0.88
NT10	77.87	84.19	81.03	0.62	0.89
CT10	80.24	79.45	79.84	0.60	0.85
NTCT10	80.63	87.75	84.19	0.69	0.91

*MCC: Matthew's correlation coefficient; AUC: Area under curve.

SVM model on peptides with length up to 10. Since the most of the THPs have length between 4 and 10, therefore, we have constructed a dataset (469 peptides) consisting of peptides having length up to 10. SVM models were developed using all the above features and terminal of window size 5. Performances of all models are summarized in Table 4. Maximum accuracy of 81.88% with MCC of 0.65 and AUC 0.88 was achieved in binary profile of dataset NTCT5 (Table 4).

ROC Plot. In order to have a threshold-independent evaluation of our models, we have generated receiver operating characteristic (ROC) curve for these models. PASW statistical package was used for creating ROC plots with area under curve (AUC). The AUC gave a single value to evaluate the performance of a method. BPP-based method in case of hybrid of N-terminal and C- terminal residues (window size 5 and 10) performed better as compared to AAC-based method. ROC plots are shown in Figure 4.

Performance on independent dataset. In order to validate our *in silico* methods, performances of our best methods (whole composition, NTCT5, NTCT10, and NTCT5 (up to 10)) were evaluated on independent dataset. All these models performed reasonably good as shown in Table 5, demonstrating that these models are useful or effective in real life. Composition-based model achieved highest accuracy of 83.73% among all these models.

Implementation and utility of TumorHPD. TumorHPD not only provides facility to predict THPs, but also offers opportunity to design analogues with better tumor homing abilities. TumorHPD first generates all possible single substitution mutants of original peptide; then it predicts whether mutants and original peptide is tumor homing or not. It also calculates SVM score for each peptide, which is propotional to reliability of prediction. Along with prediction, server also calculates important physicochemical properties (*e.g.* hydrophobicity, amphipathicity, charge, pI, *etc.*) in an aesthetic tabular format with sorting option. This feature is helpful for user to select better analogues based on desired physicochemical properties, as many peptide analogues may have higher SVM score or better-desired properties than the original peptide. In addition, users can further generate all possible mutants (2nd round) of their selected analogue if they wish to, and may get even better peptide analogues with higher tumor homing abilities (based on SVM score). This cycle can be run until the peptide analogue with desired properties (tumor homing and physicochemical) is obtained. Similarly, protein scanning is another tool, which allows user to submit protein sequence and it scan putative THPs in protein sequence. Graphical display of the scanned results speeds-up the identification of THP

Table 4 | Performance of monopeptide, dipeptide and binary profiles-based SVM models on dataset of peptides having length up to 10

INPUT VECTOR	DATASET	Sensitivity	Specificity	Accuracy	MCC	AUC
Monopeptide	Main dataset	80.81	79.74	80.28	0.61	0.87
Dipeptide	Main dataset	74.63	82.09	78.36	0.57	0.85
Monopeptide	NTCT5	78.25	80.6	79.42	0.59	0.85
Dipeptide	NTCT5	70.15	84.86	77.51	0.56	0.83
Binary	NTCT5	73.13	90.62	81.88	0.65	0.88

*MCC: Matthew's correlation coefficient; AUC: Area under curve.

specific regions from protein. In addition, users can also predict secondary structures of their peptides using Psipred[13]. TumorHPD is accessible from URL http://crdd.osdd.net/raghava/tumorhpd/.

Discussion

In the past, THPs have been successfully used as delivery vehicles to target imaging agents, drug molecules, oligonucleotides, and inorganic nanoparticles to tumors[7,10]. Most of the THPs have been identified by *in vivo* phage display technology, which is a very time consuming and laborious process. Therefore, development of an *in silico* method for predicting THPs will be very useful for biologists working in the field of peptide-based drug delivery. Thus, keeping these facts in mind, in the present study, we have made a systematic attempt to develop an *in silico* approach to predict/design THPs. The

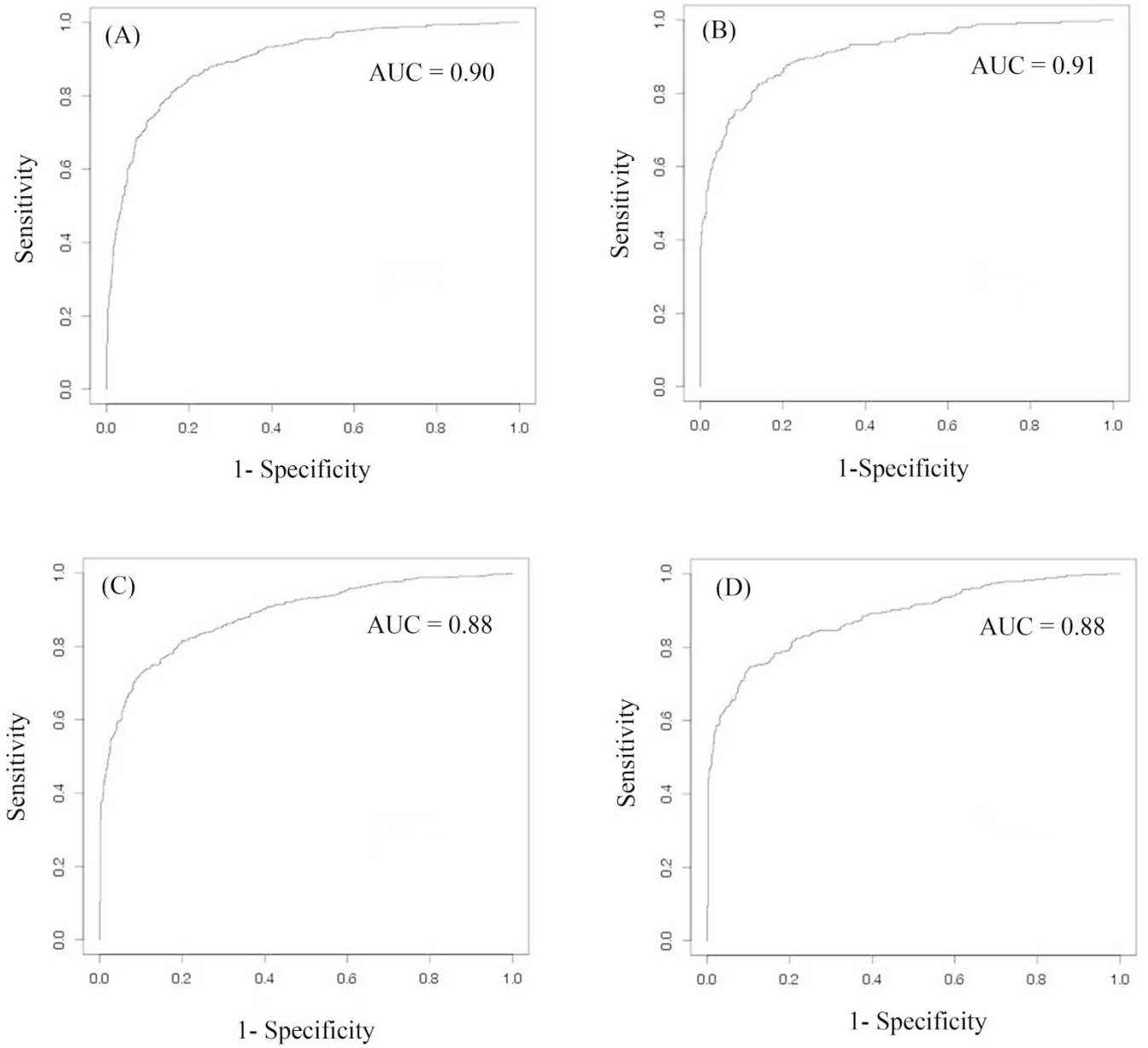

Figure 4 | ROC curves of SVM models developed using (A) Whole amino acid composition, (B) NTCT10 binary, (C) NTCT5 binary (main dataset), and (D) NTCT5 binary of dataset of peptides having length up to 10 as input features (where 1-specificity represents the false positive rate and values show the area under curve).

Table 5 | Performances on independent dataset

Model	Sensitivity	Specificity	Accuracy	MCC	AUC
Composition	77.11	90.36	83.73	0.68	0.94
NTCT5 (whole)	73.17	87.80	80.49	0.62	0.86
NTCT10 (whole)	43.28	85.07	64.18	0.31	0.76
NTCT5 (upto10)	64.18	89.55	76.87	0.56	0.82

*MCC: Matthew's correlation coefficient; AUC: Area under curve.

overall approach is summarized in Figure 5. We have collected 651 THPs from TumorHoPe database and analyzed them. THPs have wide variation in length ranging from 3 to 35 residues and majority of peptides have length between 5 and 10 residues.

In preliminary analysis of THPs, we have observed that certain residues are dominated over others and certain residues are preferred at specific positions. Based on these observations, we developed models for discriminating THPs and non-THPs using machine learning techniques. We have developed SVM models of various features using AAC, DPC and BPPs. The DPC-based models performed poorer than AAC-based method. However, BPP-based method performed well over other methods. Since binary profiles incorporate information about both frequency as well as order of amino acids, it is a better feature than AAC alone. Among all the subsets, NTCT5 and NTCT10 achieved the maximum AUC of 0.88 and 0.91 respectively. Binary performance was also best in case of peptides with length range in between 5 and 10 residues. Based on above approaches, an online web service-TumorHPD has been developed. To the best of our knowledge, TumorHPD is first *in silico* method in its kind for the prediction of THPs. Therefore, there are no existing methods for comparison. We hope that establishment of such methods will speed up the pace of identifying novel THPs. Thus, it will facilitate better drug delivery system for cancer.

Methods

Main dataset. Recently, our group has collected and compiled experimentally validated THPs (peptides bind/home to tumor) from literature and developed a public database TumorHoPe[14]. In this study, we have obtained 651 THPs from TumorHoPe. These peptides are considered as positive examples. In order to develop a classification method, we needed negative examples (*i.e.* peptides, which do not bind to tumor or non-THPs). Unfortunately, experimentally validated non-THPs have not been reported in the literature. In order to generate negative dataset, we have generated 651 random peptides from proteins obtained from SwissProt. These

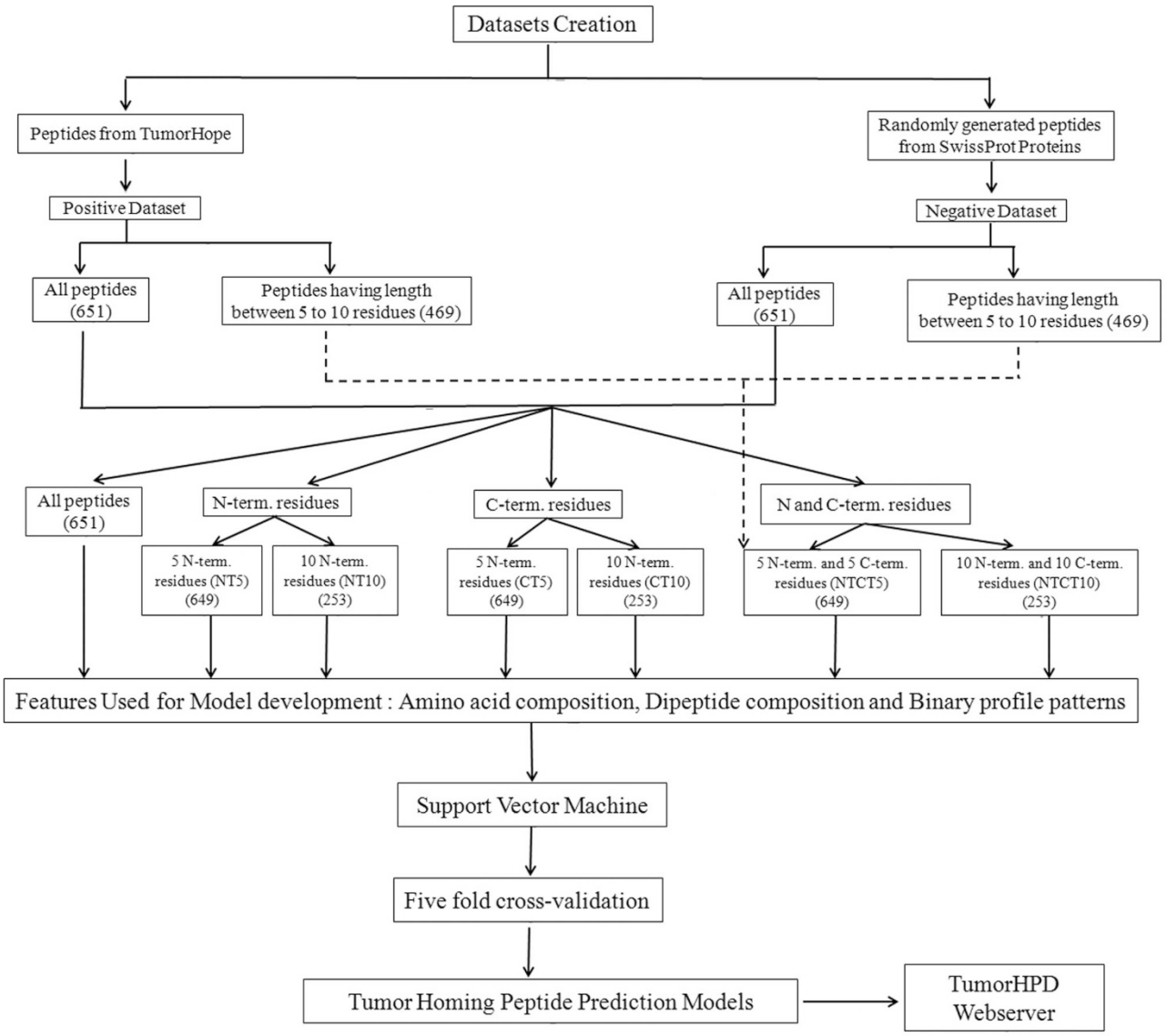

Figure 5 | Overall approach for *in silico* prediction of THPs.

random peptides were considered as non-THPs. Though it is possible that some of the random peptides may have tumor homing property, but probability is very low. This is a standard procedure to use random peptides as negative examples in situations where experimentally validated negative examples are not available[15,16]. Finally, main dataset is consists of 651 THPs (experimentally validated) and 651 non-THPs (random peptides).

Small dataset. It was observed that most of the THPs have 10 or less than 10 residues. Therfore, we created a sub dataset from main dataset where peptides (THPs or non-THPs) have minimum four residues and maximum ten residues. This small dataset contains 469 THPs and equal number of non-THPs (random peptides).

Terminus datasets. In order to understand the role of N- and C-terminal residues of THPs, we have created terminus datasets considering the N- and C-terminal residues of peptides from main dataset. Following type of terminus datasets have been derived from main dataset; (i) **NT5** contains first five residues (5 N-terminus residues) of peptides, (ii) **CT5** contains last five residues (5 C-terminus residues) of peptides, and (iii) **NTCT5:** in this dataset, various features (amino acid composition, dipeptide composition and binary profiles) of first five and last five residues of peptides were generated and combined them for developing models. Similarly, **NT10, CT10** and **NTCT10** terminus datasets were derived from main dataset where ten residues were taken either from any one terminus or from both termini.

Sequence logos. In order to understand frequency of different types on amino acids at different positions in THPs, we created sequence logos using WebLogo software[17]. The size of the residue in logo represents the frequency of residues at a given position. The height of the residue is a measure of the variability of that residue at that particular position: the taller the logo, the lesser variability at that position.

Support vector machine. SVM is a machine-learning tool based on the structural risk minimization principle of statistics learning theory. SVMs are a set of related supervised learning methods used for classification and regression. The user can choose and optimize number of parameters and kernels (e.g. Linear, polynomial, radial basis function and sigmoidal) or any user-defined kernel. In this study, we implemented SVMlight Version 6.02 package of SVM[18], which requires a fixed number of inputs for training, thus necessitating a strategy for encapsulating the global information about proteins of variable length in a fixed length format. The fixed length format was obtained from protein sequences of variable length using amino acid composition, dipeptide composition and binary profile.

Amino acid composition (AAC). It has been shown in previous studies that simple frequency of 20 amino acids in a protein sequence can be used to predict various functions of proteins like sub-cellular localization and classification of proteins[19]. In this study, we have used AAC of peptides for discriminating THPs and non-THPs. Thus, peptide information was encapsulated in a vector of 20 dimensions, using amino acid composition of the peptide. AAC is the fraction of each amino acid type within a peptide. The fractions of all 20 natural amino acids were calculated by using the following equation:

$$Comp\ \ (i)\ =\ \frac{R_i}{N}\ x\ 100$$

Where Comp (*i*) is the percent composition of amino acid (i); R_i is number of residues of type *i*, and N is the total number of residues in the peptide.

Dipeptide composition (DPC). DPC provides composition of pair of residues (*e.g.* Ala-Ala, Ala-Leu) present in peptide, and used to transform the variable length of peptides to fixed length feature vectors. It gives a fixed pattern length of 400 (20 × 20) and encapsulates information about the fraction of amino acids as well as their local order. It is calculated using following equation:

$$Fraction\ of\ Dipeptide\ (i)\ =\ \frac{Total\ number\ of\ Dipeptide\ (i)}{Total\ number\ of\ all\ possible\ dipeptides}$$

Where dipeptide (i) is one out of 400 dipeptides.

Binary profile patterns (BPP). BPP were generated for each peptide, where a vector of dimensions of 20 represents each amino acid (*e.g.* Ala by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0). A pattern of window length W was represented by a vector of dimensions 20 × W. We have created binary profile patterns for first 5 and 10 residues from N-terminus, similarly for last 5 and 10 residues from C-terminus of peptides in datasets. The BPP has been used in a number of existing methods[20–22].

Cross-validation technique. One of the major challenges in developing *in silico* models is to validate these models using standard techniques. One of the well known and commonly used technique for validation is jack-knife or leave-one-out cross-validation where one peptide is used for testing and remaining peptides for training. This process is repeated in such a way that each peptide is used for testing. This technique is CPU time intensive, so in this study we have used five-fold cross-validation technique. Here all peptides are randomly divided into five sets, where four sets used for training and remaining set for testing. The process is repeated five times in such a way that each set is used once for testing. Final performance is obtained by averaging the performance of all the five sets.

Performance measure. The performance of various models developed in this study was evaluated by using threshold-dependent as well as threshold-independent parameters. In threshold dependent parameters we used sensitivity (Sn), specificity (Sp), overall accuracy (Ac) and Matthew's correlation coefficient (MCC) using following equations.

$$Sensitivity = \frac{TP}{TP+FN}\ x\ 100$$

$$Specificity = \frac{TN}{TN+FP}\ x\ 100$$

$$Accuracy = \frac{TP\ +\ TN}{TP\ +\ FP\ +\ TN\ +\ FN}\ x\ 100$$

$$MCC = \frac{(TP)\ (TN)\ -\ (FP)\ (FN)}{\sqrt{(TP+FP)\ (TP+FN)\ (TN+FP)\ (TN+FN)}}$$

Where TP and TN are correctly predicted positive and negative examples, respectively. Similarly, FP and FN are wrongly predicted positive and negative examples, respectively.

We created receiver-operating characteristic (ROC) for all of the models in order to evaluate performance of models using threshold-independent parameters. ROC plots with area under the curve (AUC) were created using PASW statistical package.

Independent dataset. In order to evaluate the performance of our methods, we have created an independent dataset of 83 novel experimentally validated THPs and equal number of random peptides (non-THPs), which have not been included in the training, feature selection and parameters optimization of the model. Experimentally validated THPs were collected manually from recent research papers and patents, while random peptides were generated randomly from proteins obtained from Swissprot as discribed in methods.

1. Hanna, T. P. & Kangolle, A. C. Cancer control in developing countries: using health data and health services research to measure and improve access, quality and efficiency. *BMC Int Health Hum Rights* **10**, 24 (2010).
2. Lee, C., Raffaghello, L. & Longo, V. D. Starvation, detoxification, and multidrug resistance in cancer therapy. *Drug Resist Updat* **15**, 114–22 (2012).
3. Flaherty, K. T., Hodi, F. S. & Fisher, D. E. From genes to drugs: targeted strategies for melanoma. *Nat Rev Cancer* **12**, 349–61 (2012).
4. Higgins, M. J. & Baselga, J. Targeted therapies for breast cancer. *J Clin Invest* **121**, 3797–803 (2011).
5. Scott, A. M., Wolchok, J. D. & Old, L. J. Antibody therapy of cancer. *Nat Rev Cancer* **12**, 278–87 (2012).
6. Chames, P., Van Regenmortel, M., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *Br J Pharmacol* **157**, 220–33 (2009).
7. Laakkonen, P. & Vuorinen, K. Homing peptides as targeted delivery vehicles. *Integr Biol (Camb)* **2**, 326–37 (2010).
8. Zitzmann, S., Ehemann, V. & Schwab, M. Arginine-glycine-aspartic acid (RGD)-peptide binds to both tumor and tumor-endothelial cells in vivo. *Cancer Res* **62**, 5139–43 (2002).
9. Pasqualini, R. *et al.* Aminopeptidase N is a receptor for tumor-homing peptides and a target for inhibiting angiogenesis. *Cancer Res* **60**, 722–7 (2000).
10. Ruoslahti, E., Bhatia, S. N. & Sailor, M. J. Targeting of drugs and nanoparticles to tumors. *J Cell Biol* **188**, 759–68 (2010).
11. Ruoslahti, E. Peptides as targeting elements and tissue penetration devices for nanoparticles. *Adv Mater* **24**, 3747–56 (2012).
12. Petrilli, P. Classification of protein sequences by their dipeptide composition. *Comput Appl Biosci* **9**, 205–9 (1993).
13. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–5 (2000).
14. Kapoor, P. *et al.* TumorHoPe: a database of tumor homing peptides. *PLoS One* **7**, e35187 (2012).
15. Sanders, W. S., Johnston, C. I., Bridges, S. M., Burgess, S. C. & Willeford, K. O. Prediction of cell penetrating peptides by support vector machines. *PLoS Comput Biol* **7**, e1002101 (2011).
16. Wang, P. *et al.* Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS One* **6**, e18476 (2011).
17. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188–90 (2004).
18. Joachims, T. Making large-scale support vector machine learning practical. *In Advances in kernel methods: support vector learning* Edited by: Scholkopf B, Burges C, Smola A. Cambridge, MA: MIT Press, 169–184 (1999).
19. Garg, A., Bhasin, M. & Raghava, G. P. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J Biol Chem* **280**, 14427–32 (2005).

20. Xiao, X., Shao, S., Ding, Y., Huang, Z. & Chou, K. C. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. *Amino Acids* **30**, 49–54 (2006).
21. Xiao, X., Wang, P. & Chou, K. C. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. *J Comput Chem* **30**, 1414–23 (2009).
22. Lata, S., Sharma, B. K. & Raghava, G. P. Analysis and prediction of antibacterial peptides. *BMC Bioinformatics* **8**, 263 (2007).

Acknowledgements

Authors are thankful to funding agencies Council of Scientific and Industrial Research (project Open Source Drug Discovery and GENESIS BSC0121) and Department of Biotechnology (project BTISNET), Govt. of India.

Author contributions

A.S., P.K., A.G. and K.C. collected the data and created the datasets. A.S., A.T. and P.K. developed computer programs, implemented S.V.M. A.S. and J.S.C. created the back end server. A.S., P.K., A.G., R.K. and K.C. developed the front end user interface. A.G., P.K. and A.S. wrote the manuscript. G.P.S.R. conceived and coordinated the project, helped in the interpretation of data, refined the drafted manuscript and gave overall supervision to the project. All of the authors read and approved the final manuscript.

Additional information

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

How to cite this article: Sharma, A. *et al.* Computational approach for designing tumor homing peptides. *Sci. Rep.* **3**, 1607; DOI:10.1038/srep01607 (2013).