Providing Available Molecular Modeling Data for Composing Focal Adhesion Kinase Inhibitor by C 5.0, Support Vector Machine, and Structure Prediction Program

Dae Young Kim and Taeseon Yoon

***Abstract*—On the membrane of cells, for the purpose of interactions with extracellular matrix, there is a focal adhesion, which is a supramolecular assembly, consisting of various proteins. This cluster has been known for contributing to cell movement and anchorage. Especially, focal adhesion kinase, which is the most essential protein in the focal adhesion, has been detected a lot in surroundings of cancer cells which tells that the kinase is used to cancer outbreak or metastasis. A recent study, conducted by UC San Diego Moores Cancer Center, showed that in vitro, cancer cell could not penetrate or attach to endothelial cells when FAK inhibitor was applied. Under many other researches of cancer, clinical demonstration of FAK inhibitor is actively ongoing to make appropriate prevention to cancer metastasis. In this research, to find out the most available and representative of data of human FAK, we used decision tree algorithm program, C 5.0, with the 12 FAK isoforms sequence data from gene prediction program. Also, by comparing the structural differences between those two isoform with support vector machine and ligand-binding site prediction program, we suggest more accurate and effective data for production of FAK inhibitor.**

***Index Terms*—Cancer cell, focal adhesion kinase, inhibitor of focal adhesion kinase, C 5.0, support vector machine.**

I. Introduction

Extracellular matrix is composed of matrix between cells and membrane space and it is important for cells in that if cells do not have interaction with it, they would go through anoikis, the form of apoptosis. On the cell membrane, the transmembrane receptor exists which is called integrin. When the receptor binds with ligands, cell membrane alters its structures and proteins which were inside of cell matrix approach to the integrin and form supramolecular assembly. This assembly is called focal adhesion which serves a lot of function for cell including cell movement and anchorage due to various kinds of proteins in it and the protein which is the most essential in the assembly is focal adhesion kinase (FAK). According to recent research by UC San Diego Moores Cancer Center, FAK does have an crucial role in the mechanism of cancer metastasis [1].

Meanwhile, cancer outbreaks because of genetic mutation and the causes of it are so variety that not only smoke or alcohol but even excessive stress is known as the cause of cancer outbreak. In past, the exact cause of cancer and the mechanism of its metastasis were not known to researchers which made it hard to cure or prevent it. Nowadays, however, the mechanisms of cancer outbreak and metastasis are relatively well known in genetic aspect than in the past. We now know there are genes which not only generate tumors but also suppress them [2]. Those genes are encoded with the information of kinase and phosphastase which are proteins that involved in cancer outbreak. Many researches are ongoing to find out more precise and exact mechanisms of the cancer outbreak and metastasis.

As a result of being understood more about cancer, the mortality of patients with tumors has decreased while the occurrence rate has increased [3]. This shows that medical science has contributed successfully to reduce the negative symptoms and the severity of cancer.

Commonly, doctors examine their cancer patients based on the state of their cancer metastasis. The more metastasis progressed in their bodies, the higher probability of death. In other words, if cancer cells from the patient's specific organs has already been spread extensively, the patient would be diagnosed with late stage cancer and he or she would have more probability of death than early stage cancer patients. When cancer metastasis found in the patient's body, it goes through the blood or lymphatic vessels and ruins its surrounding organs which is the most hazard stage to the patient [4]. For this reason, cancer researches mainly take notice for its metastasis mechanism to prevent it.

In this research, we found the appropriate data for FAK inhibitor molecular modeling by analyzing FAK amino acid sequence and comparing the two isoforms which represent FAK best among 12 isoforms. The 12 FAK isoforms sequence data were obtained by gene prediction program of Homo Sapiens PTK 2 gene from National Center for Biotechnology Information (NCBI) [5]. We selected the two most representing isoforms from the data by using decision tree algorithm, C 5.0. Then, with these two isoforms, we picked out major sequence that classify each isoforms by using support vector machine (SVM) and found three dimensional structure template and probable-ligand binding sites of them using ligand-binding sites program (Phyre) [6], [7]. Focusing on those binding sites would be more effective for finding appropriate low molecular substance that can be bind with the sites of FAK and work as inhibitor.

Manuscript received March 17, 2014; revised May 16, 2014.

Dae Young Kim is with the Natural Science Department, Hankuk Academy of Foreign Studies, South Korea (e-mail: dae8177@naver.com).

Taeseon Yoon is with the International Department, Hankuk Academy of Foreign Studies, Korea (e-mail: tsyoon@hafs.hs.kr).

II. Genetic Factor of Cancer Outbreak and Metastasis

A. Mechanism of Cancer Outbreak

1) Cell proliferation

There are three main types of cancer outbreak processes; cell proliferation, cell immature, and programmed cell death [8]. Cells receive proliferation signals by growth factors. These factors bind with outside receptors of cells and as the signal received, phosphorus (P) transports between molecular in cell body. During this process, kinase promotes transportation of phosphorus and as the process done, phosphastase removes it from the previous molecular. At the end of this process, phosphorus is given to retinoblastoma (Rb), and then the cell proliferates (Fig. 1).

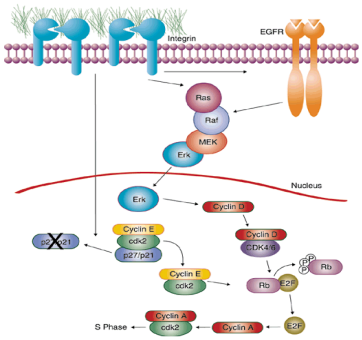

Fig. 1. Integrin and growth factor cooperation in cell cycle regulation (source: Cindy K. Miranti and Joan S. Brugge, "Sensing the environment: A historical perspective on integrin signal transduction," Nature Cell Biology, vol. 4, pp. E83-E90, 2002, Fig. 2).

When the kinase is excessively activated, however, cell proliferation would be promoted overly and the cell forms a tumor. For example, when chromosomes 9 and 22 have reciprocal translocation, each chromosome gets hybrid genes which may cause the expression of cancer genes that make kinase active. Conversely, when tumor suppressor genes are expressed, phosphastase would be active. These genes carry one cancer cell into normal and usually it slows down its proliferation speed. For example, TP53 which is on chromosome 17, is a popular tumor suppressor gene that contribute to expression of tumor suppressor protein P53 [9]. Since cancer gene and tumor suppressor gene mutually affect the cell proliferation, when such genes like TP53 goes wrong, various types of cancers can occur at once and this is called Li-Fraumeni syndrome.

2) Cell immature

The second type of cancer outbreak process is disorder of transcription factor. Transcription factor, which is a protein which controls gene expression inside of the cell, decide the cell daughter cell's function based on its state. Therefore, when there is a problem on the transcription factor, malfunctioning cell comes out from cell proliferation which can do nothing for the body but interrupts surrounding cells' operation.

3) Programmed cell death

Also, cancer may be outbroke by the disorder of Bcl2. When cell expired its age, it goes through a programmed cell death (apoptosis) [10]. In the process, Bcl2 opens the hole of mitochondria surface, causing substances inside of the cell to come out and DNA to be cut, making the cell die. However, if the gatekeeper of mitochondria surface, Bcl2, cannot open the hole because of some stimulation, the old cell will never die, staying unusable and this becomes a tumor cell.

B. Cancer Metastasis

Some of the cancer cells secrete vascular endothelial growth factor (VEGF) which makes blood vessel cells around the cancer to proliferate and grow toward the tumor and allows it to get essential substances, nutrients and oxygen from the proliferated vessels for its survival. Even if those vessels which were grown around tumor are not as normal and functional as original vessels, when the tumor is supplied nutrients from them, they grow much faster than ever. In addition, when it grows its scale enough size to approach near vessels, the cancer cells make an opening to their organ cells' wall and penetrate into blood or lymphatic vessels. Going through a series of mechanism, the cells metastasis throughout the body and the spread cancer cells go to another organs and cause trouble (Fig. 2.)

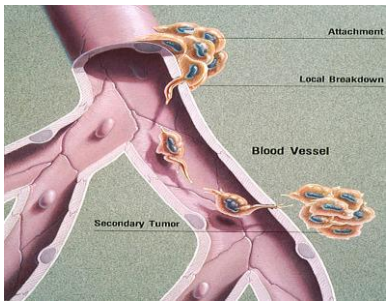

Fig. 2. Metastasis illustration; How cancer spreads (source : National Cancer Institute, Jane Hurd).

Actually, not all cancer cells have the feature of metastasis; only a few of them emit VEGF. However, once the vessels grow to the tumor, the process of metastasis starts by penetrating into blood and lymphatic vessels. Even if our human immune system gets rid of almost all of the cancer cells that are spreading through blood vessels, a few of them occasionally hide behind fibrin membrane and evade the attack of immune system [11]. In lymphatic vessels, sentinel lymph node filter out cancer cells that goes through the vessels, but when the cells come so many that the filtering effect reaches limit, cancer cells penetrates the nodes and metastasis through the body. When these survived cancer cells reached to capillary vessels around other organs, they stuck in there which is called tumor embolism. Then, the cells spread around the surrounding cells to find the appropriate receptors for their anchorage. After they attached to other cells, they become another tumor and cause problems at the region.

III. Focal Adhesion Kinase

A. What Is FAK?

Extracellular matrix consists of matrix between cells and basement membrane. This matrix plays a crucial role in cell survival. It interacts with cells, supports them and maintain structures. Therefore, if cells are located far from extracellular matrix and cannot interact with it, the cells would die [12]. This death is called anoikis which is a kind of apoptosis, a programmed cell death. Integrin is a transmembrane receptor which distributes on the surface of cell membrane and it mediate the attachment between cells

and extracellular matrix. When integrin forms binding with ligand, a minute change of membrane occurs and various proteins from inside of cell membrane approach to form a supramolecular assembly which is called focal adhesion. In the assembly, those various kinds of proteins such as Grb2, GRAF are assembled which makes focal adhesion to perform various kinds of function including cell movement and anchorage. FAK, which is one of these proteins, performs the most essential role among them such as providing binding sites for other proteins [13]. It is encoded in PTK2 gene which is in chromosome 8 and is also called protein tyrosine kinase 2 because it provides binding sites by phosphorylating its tyrosine (Y397). The cell which has many FAK can evade anoikis process without interaction with extracellular matrix because FAK suppress anoikis by its enzymatic function and phosphorylation of Y397. Therefore, FAK is crucial for cancer cells survival since cancer cells cannot interact with extracellular matrix during metastasis.

B. FAK Inhibitor

FAK performs two kinds of functions; enzymatic function and proteomic function. The most common enzymatic function of FAK is contributing to survival of cancer cell as mentioned in III. *Focal Adhesion Kinase.* It prevents cancer cells from going through the anoikis by phosphorylation of Y397. On the other hand, FAK also plays a crucial role, which is a proteomic function, in penetrating and anchorage of cancer cells to endothelial cells [14].

Therefore, if we can inhibit FAK's action for assisting cancer cells' survival, it would be easy to stop cancer metastasis through patient's body and to do so, understanding FAK proteins' interactions among other proteins is important. Therefore, to figure out interactions among protein sites, the FAK's overall structure and ligand- binding sites should be figured out first. And then, by using simulation program such as high throughput screening, we should find proper low-molecular weight substances which can bind well enough to interrupt FAK's interactions. However, this method needs quite precise measurement and data because finding the proper substances that bind to FAK is hard. Also, the research should be conducted not only in vitro, but also in vivo to examine exactly for clinical usage [15].

Therefore, to provide more efficient aspect of production of FAK inhibitor in clinical usage, we used decision tree algorithm, C 5.0, to classify the data of 12 Homo Sapiens FAK amino acid isoform sequences which were generated from gene prediction program by using PTK 2 gene in chromosome 8. After classifying two isoforms as the most representing class, we also found different sites between them by using SVM with Kernel methods. Also, we used three-dimensional protein structure prediction program (Phyre) which is based on ab inito method and found out the most accurate template structure [6].

IV. Experiments

Step 1: Decision Tree Algorithm (C 5.0)

Above all, we obtained 12 isoforms of FAK amino acid sequence data from PTK2 gene in chromosome 8 using genomic sequence prediction program from NCBI. [5] Then, we needed to select particular isoform data which shows the highest represent ability of FAK. In this process, we used decision tree algorithm, C 5.0. Decision tree algorithm, which is one of the most typical methods for data mining, can perform a function of classifying given data and finding classes that show represent ability. We conducted experiment with given sequence data of 12 isoforms in window size 9, 13, and 17 respectively and used 10 fold cross validation method. Then, for each class, we extracted overlapping rule sets, whose frequency values are over 0.75. Table I, Table II, and Table III are extracted rule sets for each class and their frequency values.

TABLE I: Rule Set Results from C 5.0 in Window Size 9

Class Number	Amino Acid Positions	Frequency Average
4	pos3 = **V** pos4 = **R**	0.750
	pos2 = **P** pos4 = **G**	0.750
10	pos4 = **P** pos7 = **F**	0.750
	pos3 = **H** pos4 = **K**	0.750
11	pos1 = **N** pos2 = **L**	0.750
	pos6 = **Y** pos7 = **Q**	0.750
	pos3 = **S** pos5 = **G** pos6 = **S**	0.750
	pos4 = **R** pos6 = **F** pos7 = **L**	0.750

TABLE II: Rule Set Results from C 5.0 in Window Size 13

Class Number	Amino Acid Positions	Frequency Average
3	pos11 = **Q** pos12 = **E**	0.750
	pos3 = **Q** pos10 = **G**	0.750
	pos3 = **K** pos10 = **G**	0.750
4	pos6 = **A** pos10 = **L**	0.750
6	pos4 = **P** pos10 = **R**	0.750
7	pos8 = **D** pos10 = **P**	0.750
	pos2 = **M** pos10 = **P**	0.750
	pos10 = **A** pos12 = **A**	0.750
8	pos8 = **T** pos10 = **P**	0.750
10	pos4 = **M** pos10 = **R**	0.750
	pos10 = **H** pos12 = **V**	0.750
	pos3 = **M** pos10 = **G**	0.750
	pos8 = **D** pos10 = **H**	0.750
11	pos2 = **V** pos10 = **P**	0.750
	pos10 = **P** pos13 = **T**	0.750
12	pos3 = **D** pos10 = **F**	0.750

For each window size, order of classes which have the best rule sets is shown Table IV.

After aligning the order, we valued each rank for certain numerical value; +3 to 1st, +2 to 2nd, and +1 to 3rd. After summing the numerical values of each class in whole window sizes, we found out that class 7 and 10 are the most representing isoform sequence data for FAK.

Step 2: Support Vector Machine

To confirm the difference between isoform 7 and 10 which show the highest represent ability in their 12 classes groups, we used support vector machine, one of the machine learning methods. First, we changed isoform 7 and 10 sequence data into number data format as Table V.

Then, we conducted experiment with linear and non-linear procedure respectively in window size 9, 13, and 17. For

linear separation of data, we used normal function and linear polynomial function and for non-linear separation, we used Radial Basis Function and quadratic polynomial function as kernel function. As shown in Table VI, precision values (accuracy values) are higher when using Radial Basis Function or quadratic polynomial function than using the others.

TABLE III: RULE SET RESULTS FROM C 5.0 IN WINDOW SIZE 17

Class Number	Amino Acid Positions	Frequency Average
4	pos2 = **P** pos5 = **I**	0.75
6	pos5 = **H** pos7 = **V**	0.75
	pos8 = **K** pos16 = **N**	0.75
7	pos5 = **L** pos14 = **N**	0.75
	pos5 = **P** pos13 = **S**	0.75
	pos1 = **D** pos5 = **S**	0.75
	pos3 = **T** pos5 = **A**	0.811
	pos8 = **E** pos13 = **P**	0.75
10	pos1 = **M** pos5 = **S**	0.7875
	pos6 = **I** pos7 = **A**	0.75
	pos8 = **Q** pos17 = **L**	0.75
	pos8 = **K** pos16 = **G**	0.75
11	pos5 = **M** pos16 = **M**	0.75
	pos3 = **A** pos5 = **A**	0.75
	pos5 = **K** pos13 = **N**	0.775
	pos8 = **K** pos16 = **S**	0.75
12	pos5 = **V** pos9 = **S**	0.75

TABLE IV: CLASS NUMBERS WHICH HAVE HIGH FREQUENCY VALUE OF RULE SETS FOR EACH WINDOW SIZES

	1st	2nd	3rd
Window Size 9	11	4 = 10	
Window Size 13	10	3 = 7	
Window Size 17	7	10	11

TABLE V: NUMBER FORMAT FOR AMINO ACID SEQUENCE

Amino Acid	Number Representation
Alanine (A)	1
Cysteine (C)	2
Aspartate (D)	3
Glutamate (E)	4
Phenylalanine (F)	5
Glycine (G)	6
Histidine (H)	7
Isoleucine (I)	8
Lysine (K)	9
Leucine (L)	10
Methionine (M)	11
Asparagines (N)	12
Proline (P)	13
Glutamine (Q)	14
Arginine (R)	15
Serine (S)	16
Threonine (T)	17
Selenocysteine(U)	18
Valine (V)	19
Tryptophan (W)	20
Tyrosine (Y)	21

Based on these results, we picked out support vector sequence from the results of Radial Basis Function and quadratic polynomial function which showed higher precision value. After designating class 7 as "+1", and class 10 as "-1", we found 4 support vector sequences that appeared the most frequently in each regions (see Table VII).

TABLE IV: PRECISION VALUE OF EACH WINDOW SIZE 9, 13, AND 17 FOR VARIOUS FUNCTIONS

	Window Size 9	Window Size 13	Window Size 17
Normal Function	53.33 %	33.33 %	64.71 %
Linear Polynomial Function	53.33 %	44.44 %	55.00 %
Quadratic Polynomial Function	70.00 %	81.25 %	78.95 %
Radial Basis Function	88.89 %	65.22 %	68.18 %

TABLE VII: SUPPORT VECTOR SEQUENCE OF EACH ISOFORMS; CAUSING DIFFERENCES IN STRUCTURAL FUNCTIONS BETWEEN THEM

Isoform	Support Vector Sequence
7	G - D - E - T - D - D - Y - A - E
	H - G - V - K - P - F - Q - G - V
	E - P - T - T - W - A - S - I - I - R - H - G - D
	V - D - M - G - U - S - S - V - R - E - K - Y – E
10	D - E - A - R - D - Y - E - I - Q
	F - T - S - A - S - D - V - W - M
	A - Y - Q - L - S - T - A – L - A
	R - F - L - P - L - V - F - C - S

Step 3: Ligand-Binding Site Prediction (Phyre)

Using sequences that decide the critical difference between isoform 7 and 10 from support vector machine and applying ligand-binding prediction program, Phyre2, we apprehended the most realistic (credible) ligand-binding sites which determine structural features and functions of FAK. Although those two isoforms' structures were predicted by similar template, where there exist some differences between them. (Fig. 3) Amino acid residue numbers were different and even if main ligand-binding sites are similar, considering that some ligand-binding sites exist in other clusters, difference between those two isoforms should be considered in molecular modeling for FAK inhibitor.

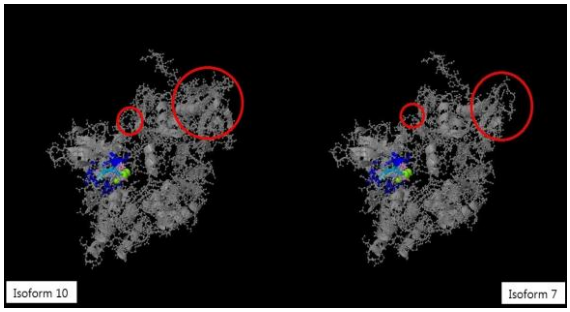

Fig. 3. The 3D structure of FAK isoform 7 and 10; Blue spheres are amino acid binding sites. Red circled points are relatively evident parts that shows differences between those two isoforms.

V. CONCLUSION AND FURTHER RESEARCH DIRECTION

To conclude, this research mainly focused on the function of focal adhesion kinase which helps cancer metastasis and analyzed the protein's relatively exact structural feature from various isoforms and found out the difference between two representing isoforms. Especially, in focal adhesion kinase inhibitor production procedure, we provided proper data of

focal adhesion kinase for molecular modeling. However, the important part of the inhibitor research is not in vitro condition but in vivo. Even if we understand all of the focal adhesion kinase's ligand-binding sites, we cannot assure the effect of inhibitor produced from the data because of some variations in vivo condition. Finding out some not only exact ligand-binding sites of focal adhesion kinase but also entropy in interactions may allow more accurate understanding and prevention of cancer metastasis.

REFERENCES

[1] C. Jean *et al*., "Inhibition of endothelial FAK activity prevents tumor metastasis by enhancing barrier function," *The Journal of Cell Biology*, Dec. 9th, 2013

[2] L. Degos, *Peut-on Vaincre le Cancer*? 2004, pp. 31-36.

[3] R. Siegel, D. Naishadham, and A. Jemal, *Cancer Statistics 2012*, pp. 10-11.

[4] J. S. Yeo, "Nuclear imaging of cellular proliferation, department of nuclear medicine," Asan Medical Center University of Ulsan College of Medicine, Seoul, Korea, vol. 38, 2004, p. 198.

[5] Gnomon supported by mRNA and EST evidence, Automated Computational Analysis, from genomic sequence NT_008046.17, NCBI Reference Sequence, XP_005251062.1.

[6] L. A. Kelley and M. J. E. Sternberg, "Protein structure prediction on the web: A case study using the Phyre server," *Nature Protocols 4*, pp. 363-371, 2009.

[7] M. N. Wass, L. A. Kelley, and M. J. Sternberg, "3DLigandSite: predicting ligand-binding sites using similar structures," *NAR 38*, pp. W469-W473, 2010.

[8] L. Degos, *Peut-on Vaincre le Cancer*? 2004, pp. 16-28.

[9] G. M. Cooper, *Oncogenes*, 1995, chapter 10, pp. 145-155.

[10] G. M. Cooper, *Oncogenes*, 1995, chapter 19, pp. 323-328.

[11] M. Reitz, *Die Chaos-Zellen*, 2006, pp. 281-302.

[12] Y. S. Lee, "The role of focal-adhesion kinase in Cancer - A new therapeutic opportunity," *1999-2005 KOSEN*, pp. 1-3.

[13] E. Zamir and B. Geiger, "Molecular complexity and dynamics of cell–matrix adhesions," *Journal of Cell Science*, vol. 114, no. 20, pp. 3583–90, 2001.

[14] D. Riveline, E. Zamir, N. Q. Balaban, U. S. Schwarz, T. Ishizaki, S. Narumiya, Z. Kam, B. Geiger, and A. D. Bershadsky, "Focal contacts as mechanosensors: externally applied local mechanical force induces growth of focal contacts by an mDia1-dependent and ROCK-independent mechanism," *Journal of Cell Biology*, vol. 153, no. 6, pp. 1175–862001.

[15] M. R. Arkin and J. A. Wells, "Small molecule inhibitors of protein-protein interactions: progressing towards the dream," *Nat. Rev. Drug Discov*., vol. 3, pp. 301-317, 2004.

Dae Young Kim was born in Daegu, South Korea in 1996. He is now in Hankuk Academy of Foreign Studies. He is interested in influenza virus research and published papers about H5N1 hemagglutinin sequence analysis with artificial neural network on International Conference on Engineering and Applied, Science. He also submitted a paper about comparison of several influenza a viruses' glycoprotein and accepted to BDM 14. He is now majoring bioinformatics, and proteomics for molecular modeling in the academy.

Taeseon Yoon was born in Seoul, Korea, in 1972. He got the Ph.D. candidate degree in computer education from the Korea University, Seoul, Korea, in 2003.

From 1998 to 2003, he was with EJB analyst and SCJP. From 2003 to 2004, he joined the Department of Computer Education, University of Korea, as a lecturer and Ansan University, as an adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a computer science and statistics teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.