

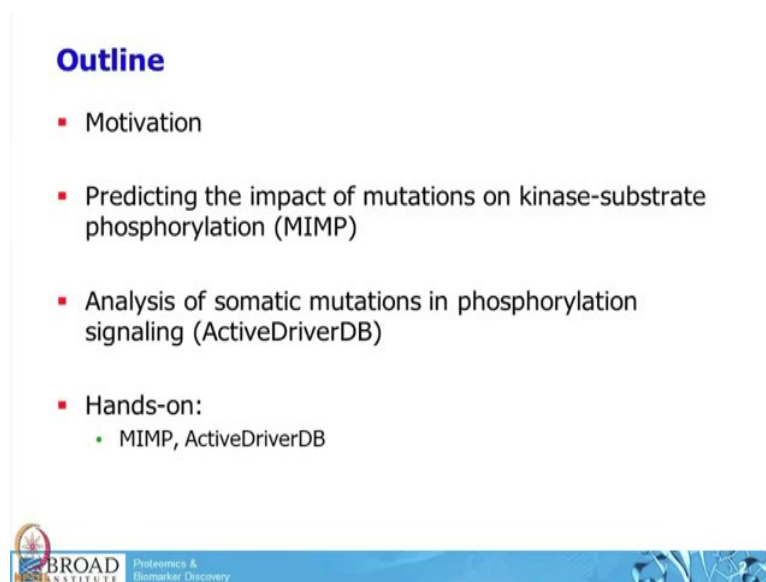
Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. Karsten Krug
Department of Biosciences and Bioengineering
Broad Institute of MIT and Harvard
Indian Institute of Technology, Bombay

Lecture – 53
Mutation and Signaling – 1

Welcome to MOOC course, on Introduction to Proteogenomics. Our next speaker is Dr. Karsten Krug, who will talk about Effect of Mutations on Signaling Pathways and how they could be studied using software's like MIMP and ActiveDriverBD. He will talk about the frequency of phosphorylation and factors, which may lead to a specific kinase activity. He will also talk about tools like motif-X and phosphosite plus for sequence motif analysis and also calculate the frequency of most recurring amino acids near the site of phosphorylation.


So, let us now welcome Dr. Karsten Krug, to talk in more detail about the rule of various mutations on signaling pathways and also to tell us about various factors which may help us in understanding how phosphorylation can be understood in a biological system.

(Refer Slide Time: 01:32)



Outline

- Motivation
- Predicting the impact of mutations on kinase-substrate phosphorylation (MIMP)
- Analysis of somatic mutations in phosphorylation signaling (ActiveDriverDB)
- Hands-on:
 - MIMP, ActiveDriverDB

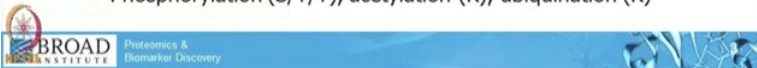
 **BROAD INSTITUTE** Proteomics & Biomarker Discovery

So, I will first going to give you like a short motivation. So, what this is why I want to do that and then I will be very specific and I will talk about two specific software tools that tried to you know that tried to study the impact of mutations on phosphorylation networks.

(Refer Slide Time: 01:46)

Genotype-to-phenotype associations unknown for the majority of mutations

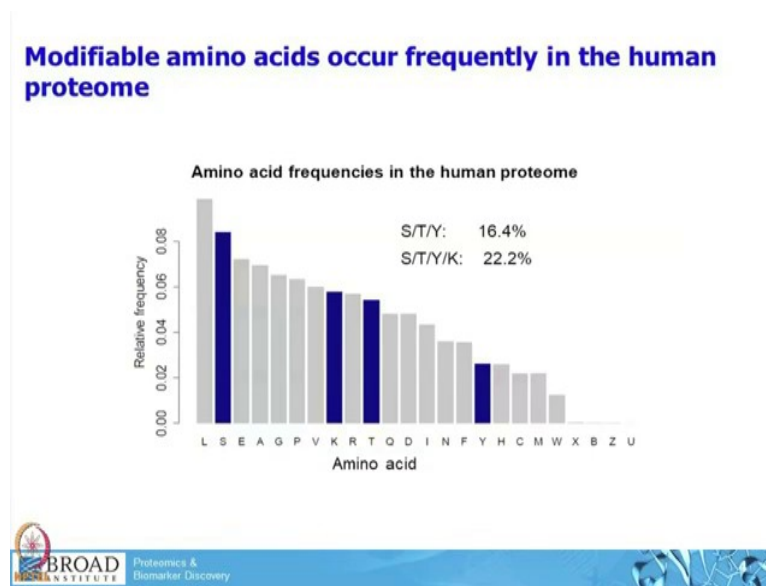
- Millions of single nucleotide variants (SNVs) in human genomes are known and thousands are associated with disease.
- Mutations located in protein coding regions can be non-synonymous (ns) and lead to single amino acid substitutions (SAV)
 - nsSNV, SAV, SAAV
- Many mutations affect post-translational modification (PTM) sites.
 - Phosphorylation (S/T/Y), acetylation (K), ubiquitination (K)



So, as Bing just presented, so, there is millions of single nucleotide variants known in the human genome and many of them are associated with certain human diseases, but for most of them we do not know the exact molecular mechanism that you know causes this genotype to phenotype association. And, as we have learned yesterday in during Kelly's and David's talk and also to the hands on so, mutations can if they are located in protein coding regions, they can be non-synonymous meaning they can lead to a single amino acid substitution. So, they can change an amino acid in the protein sequence.

And, here so, I mean there is many different synonyms for these events and non synonymous as well as single amino acid variants or single amino acid variants. So, this all you know we thought is the same kind of event, but you will find all of these and in literature. And, of course, many of these single nucleotide or non-synonymous single nucleotide variants they affect sites or amino acids that can be post-translationally modified like phosphorylation, acetylation or ubiquitination.

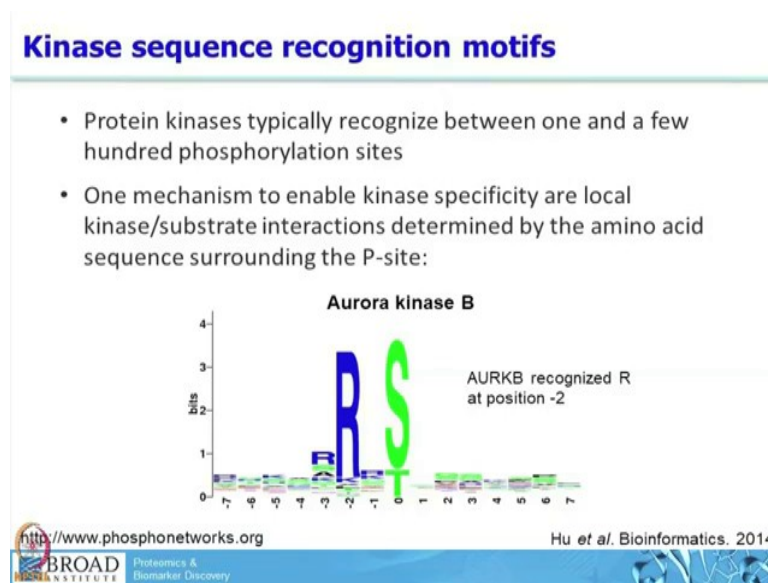
(Refer Slide Time: 03:02)



And, actually these modifiable amino acids occur very frequently in that human genome. So, what you are looking at here are the frequencies of all 20 amino acids in the human proteome. So, the most frequent amino acid that that occurs is leucine but then on the second place we already find serine which can be phosphorylated right. What's what are highlighted here are serine lysines threonine and tyrosines so, lysines or you know can be modified by acetylation or ubiquitination.

So, these are the most well studied and you know. So, we have the technology to study these modifications on a large scale. So, that is why I highlight them the highlighted though is amino acids here. And, if you just look at the like you know the overall frequencies of these 4 amino acids they make up for 22 percent of all amino acids that are occur in human proteome. So, it is very likely that a mutation affects these modification sites and we are asking the question what kind of consequences does that imply in downstream signaling events.

(Refer Slide Time: 04:17)



And, this has been many studies out there they try to you know decipher these kind of relationships and I am just highlighting a couple of those here. So, you know the most basic kind of approach to take here is to look at kinase sequence recognition motives. So, kinase phosphorylates its substrate and one mechanism to ensure that the kinase you know specifically identifies its substrate is you know is given by a by local interactions meaning it is amino acid sequence the properties of the amino acid sequence around the phosphorylation site.

So, basically as we probably many of you have seen these kinds of sequence logo motifs here, where in the center you are looking at the actual modification sites or this Aurora kinase B mostly phosphorylated serines, but also threonines and if you look at the frequency around its substrates you know flanking this sequence you see that there is a strong enrichment of an arginine at position minus 2. We got you know relative to the phosphorylation site. So, this is what we call the sequence with this motif overall kinase B. So, basically it recognizes the arginine at minus 2. Any questions to that?

Student: Sir, my question is that on what basis we are taking serine beyond in the center with 0 position?

Well, I mean our kinase I mean there is two classes of kinase. So, one separate class is tyrosine kinases which only specifically phosphorylated tyrosines. If you look at all known substrate of our kinase, we find most of them have a I mean most sites are serine sites and

then there is a smaller fraction of threonine sites. So, our kinase cannot phosphorylate tyrosines.

Student: Ok. Sir, my question was that why did they we centered at the 0.

So, we are looking solution.

Student: We have arginine also. So, like why we have to take them that as a 0 position.

So, the 0 position is actually the site where the phosphorylation happens.

Student: Ok.

And we are looking around this phosphorylation left and right.

(Refer Slide Time: 06:58)

Principle of sequence motif analysis

- Test for enrichment of amino acid patterns surrounding phosphorylation sites by comparing against a background dataset
- Phosphorylation site data set:
 - All detected sites
 - Down regulated p-sites after kinase inhibition
 - ...
- Background dataset:
 - All detected P-sites
 - All known P-sites

Phosphorylation data set	Background data set
EVVAKKAAVLLSDREKAAK	EVVAKKAAVLLSDREKAAK
EEQQAQRECVVEEAKKQTH	EEQQAQRECVVEEAKKQTH
FRKLAEKRAIPGVVTSDDQ	FRKLAEKRAIPGVVTSDDQ
FPQAFPEREVEEEDVQVGA	FPQAFPEREVEEEDVQVGA
FGDELLQDPPFFSALTQVGA	FGDELLQDPPFFSALTQVGA
THREKILLVDPQVNDQDQ	THREKILLVDPQVNDQDQ
DDMPSEKAKITPQELSEEP	DDMPSEKAKITPQELSEEP
FPFELMCKLQFRRKQSEEE	FPFELMCKLQFRRKQSEEE
VVLLTPVYIIIGASVKTAVS	VVLLTPVYIIIGASVKTAVS
GLEPYLEQELIATYVYVYV	GLEPYLEQELIATYVYVYV
EQQETKTRPFDLLEKVAAD	EQQETKTRPFDLLEKVAAD
TLKQRRRGGVYKGGKLEP	TLKQRRRGGVYKGGKLEP
DTATQGVNIGQGFQKLEET	DTATQGVNIGQGFQKLEET
EVVAKKAAVLLSDREKAAK	EVVAKKAAVLLSDREKAAK
EEQQAQRECVVEEAKKQTH	EEQQAQRECVVEEAKKQTH
FRKLAEKRAIPGVVTSDDQ	FRKLAEKRAIPGVVTSDDQ
FPQAFPEREVEEEDVQVGA	FPQAFPEREVEEEDVQVGA
EVVAKKAAVLLSDREKAAK	EVVAKKAAVLLSDREKAAK
FRKLAEKRAIPGVVTSDDQ	FRKLAEKRAIPGVVTSDDQ
FPQAFPEREVEEEDVQVGA	FPQAFPEREVEEEDVQVGA
FGDELLQDPPFFSALTQVGA	FGDELLQDPPFFSALTQVGA
THREKILLVDPQVNDQDQ	THREKILLVDPQVNDQDQ
DDMPSEKAKITPQELSEEP	DDMPSEKAKITPQELSEEP
FPFELMCKLQFRRKQSEEE	FPFELMCKLQFRRKQSEEE
VVLLTPVYIIIGASVKTAVS	VVLLTPVYIIIGASVKTAVS
GLEPYLEQELIATYVYVYV	GLEPYLEQELIATYVYVYV
EQQETKTRPFDLLEKVAAD	EQQETKTRPFDLLEKVAAD
TLKQRRRGGVYKGGKLEP	TLKQRRRGGVYKGGKLEP
DTATQGVNIGQGFQKLEET	DTATQGVNIGQGFQKLEET
EVVAKKAAVLLSDREKAAK	EVVAKKAAVLLSDREKAAK
EEQQAQRECVVEEAKKQTH	EEQQAQRECVVEEAKKQTH
FRKLAEKRAIPGVVTSDDQ	FRKLAEKRAIPGVVTSDDQ
FPQAFPEREVEEEDVQVGA	FPQAFPEREVEEEDVQVGA

BROAD INSTITUTE Proteomics & Biomarker Discovery

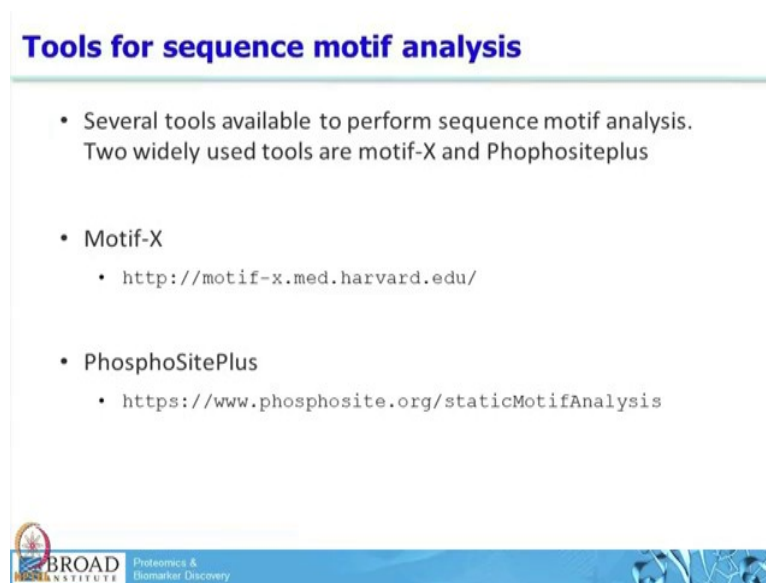
So, and there is again there is many tools that can generate these kind of logos and perform these enrichment tools. And, you know the common principle of these tools is to test for enrichment of amino acid patterns that surround this phosphorylation site and you compare it against some background dataset.

So, for example, you have your phosphorylation site data set that that you have required in your lab you know you have like in this case maybe like 28 and sites or so. So, these can be all of you all detected sites in your experiment or if you let us say you specifically inhibited the kinase and now you are looking for all phosphorylation sites that are down regulated upon

in inhibition. So, this could tell you these are very likely substrates either direct or indirect of this particular kinase.


And, then you compare, so, you compare the frequencies that you obtain here against background data set and this can again be all detected phosphatases in your data set or you could use all known phosphorylation sites in human proteome for example.

(Refer Slide Time: 08:00)



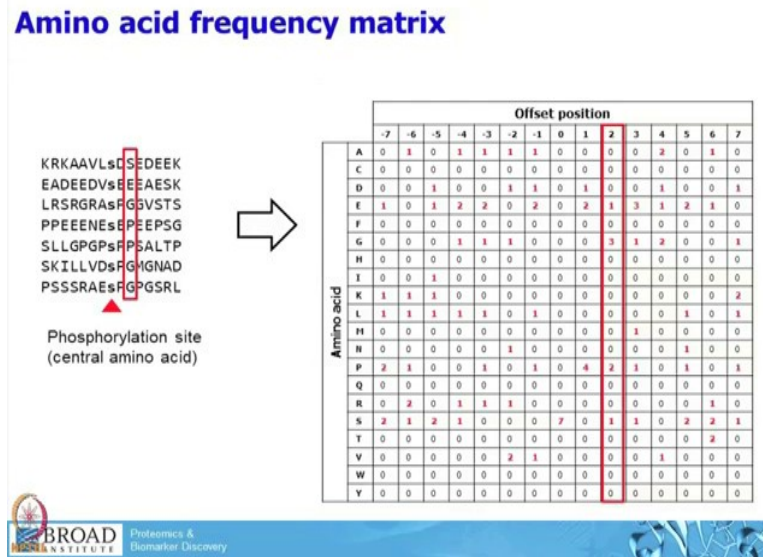
Tools for sequence motif analysis

- Several tools available to perform sequence motif analysis. Two widely used tools are motif-X and Phosphositeplus
- Motif-X
 - <http://motif-x.med.harvard.edu/>
- PhosphoSitePlus
 - <https://www.phosphosite.org/staticMotifAnalysis>

 **BROAD** INSTITUTE Proteomics & Biomarker Discovery

So, as I mentioned just several tools. So, two very popular ones are motif-X and the sequence motif analysis tool on phosphosite plus. And, motif-X was probably one of the first if not the first tool which was published back in 2005 by Steven Gygi's lab at Harvard Medical School.

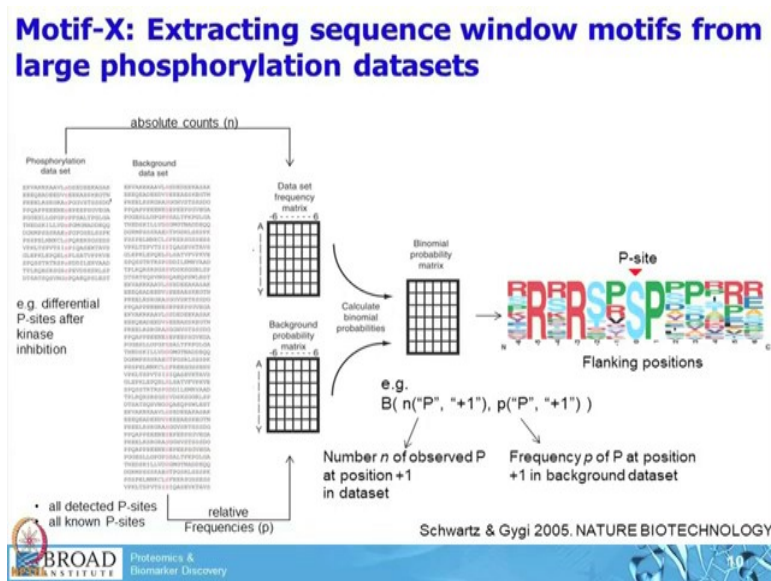
(Refer Slide Time: 08:23)



And, so, the basic principle again, so, you have your phosphorylation sites. So, the sender is where the phosphorylation happened, then we are looking at the surrounding amino acids and from that you can they easily build up this kind of frequency matrix where in your columns you have your offset positions. Again, in this in the like 0 means these are all my actual phosphorylation site and then you are looking 7 amino acids to the left and 7 amino acids to the right in this case.

Again, this is very arbitrary. So, some tools use different you know sequence windows lengths and so on. And, on the y-axis you are looking at the like all 20 amino acids. And, then you have just basically count to frequencies of these amino acids in your data. For example, in this case we would have like 3 glycines exactly we would have 2 prolines and so on and so forth right. So, it is very easy to build up this kind of frequency matrix.

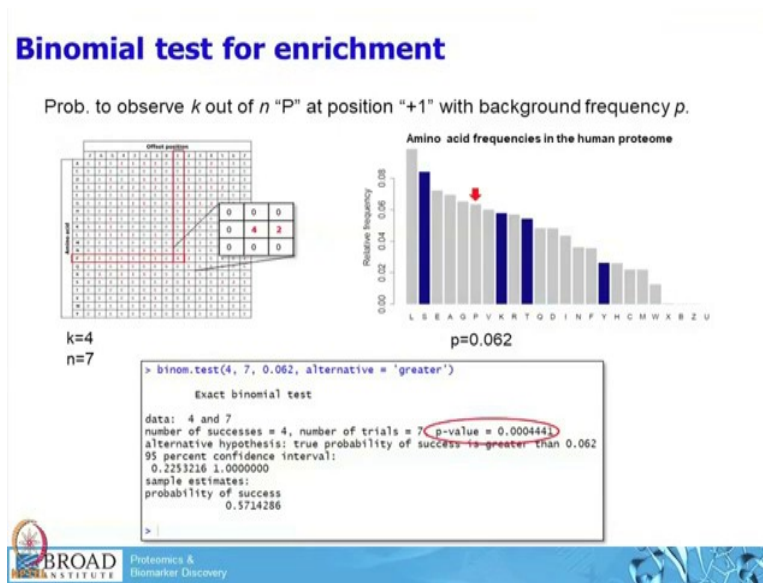
(Refer Slide Time: 09:28)



And, this is exactly what motif-X starts with builds up these two matrices – one is divided from your actual phospho-site of interest and the other is matrix is divided from your background data set. And, from that you can then calculate a binomial matrix a binomial probability matrix where you for each position in your sequence window you ask and for each amino acid you ask the question.

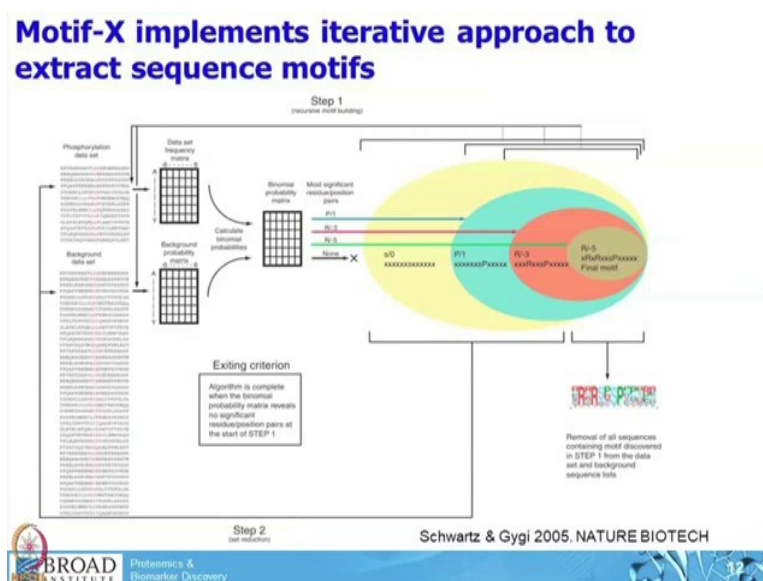
For example; in this case how many times do you observe a proline at position plus 1 in the dataset. This you can calculate for each amino acid in each position and you compare it against the background frequency which you derive from your second dataset which can again be the entire human proteome or your or all of your detected sites and from that you can then calculate or generate the sequence windows.

(Refer Slide Time: 10:25)



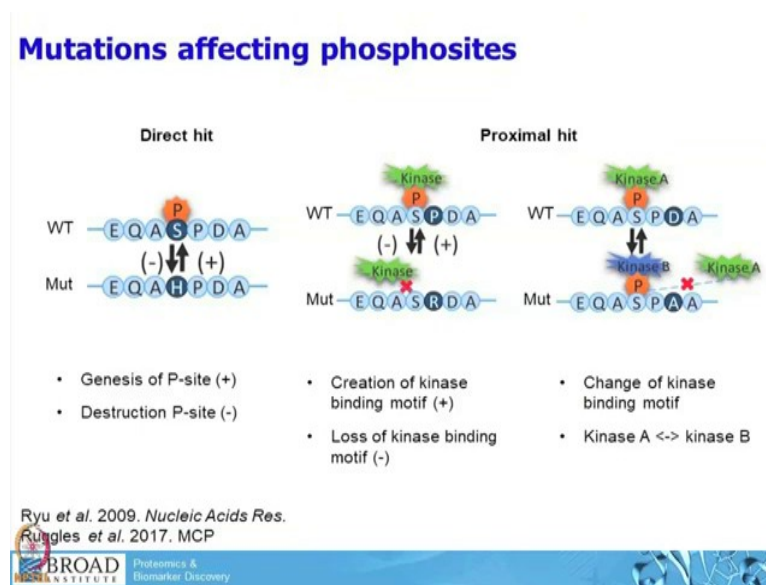
Yeah, just put up an example which is based on the 7 sequence windows that that I that we have looked at a couple of slides earlier. So, if you actually calculate this probability to observe k out of n . So, n in this case were 7. So, we looked at 7 sequence windows we observe 4 prolines at position plus 1 you know in the background probability of a proline in the human proteome is roughly 0.062 and you know in R you can just feed it into the binom test function and you get a P-value that this indeed although it is very small sample size it would be statistically significant, so, 4 out of 7 so on so forth. This again you would do for all amino acids in all occasions.

(Refer Slide Time: 11:13)



So, or you do not have to go too much into detail here is the motif-X dusted like in an iterative manner. So, it first of all takes all sites that you feed in and extracts the most significant sequence motif from that set and then if we moves those from the initial set and repeats the analysis. So, that is one as you know specific property of the software ok.

(Refer Slide Time: 11:43)



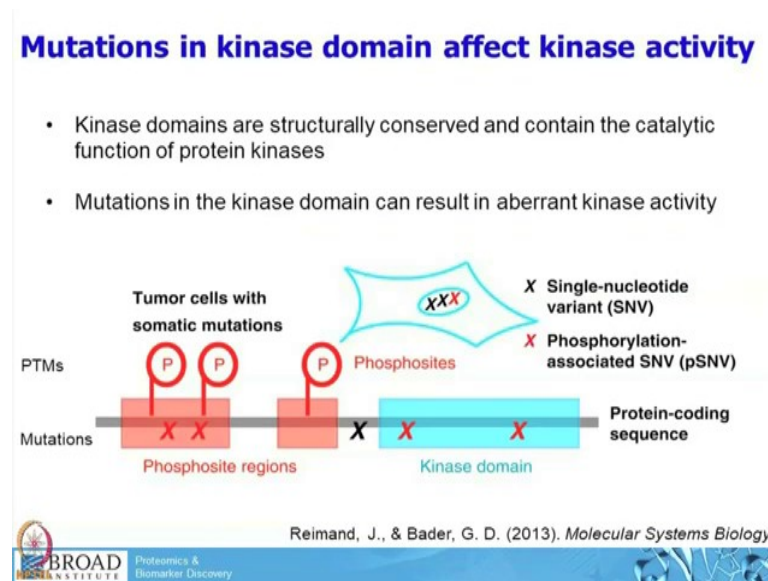
So, now we know how we can look or how we can determine the specific kinase of sequence motifs, but now what happens if these if a mutation happens in this kind of region around the phosphorylation side. So, there is actually three different scenarios that can happens. So, one is a direct hit.

So, you actually so, this is the wild type here this is a mutated version. So, wild type you have the serine which is actually phosphorylated and the serine now due to a mutation gets mutated into a histidine. So, it is cannot be phosphorylated anymore and of course, this can also happen the other way around right. So, the histidine can be mutated into a serine which now may be present of a new phosphorylation site which can be recognized by kinase.

So, it can be either like you can lead to a genesis or phosphorylation site or to a distruction of phosphorylation site. So, the other possibility that could happen so, it does not happen at the exact site, but it can happen very close to the phosphorylation site. So, in this case we have this proline here at plus 1 which now is mutated into arginine and the kinase that was able to recognize this proline can now not phosphorylate this specific serine anymore because the proline is gone right.

So, meaning in this case we would lose this phosphorylation site. Again, it can also go the other way around or it can also happen that you just change to sequence kinase motif. So, in this case and the wild type it was kinase A who recognized this motif and now due to due to a mutation this motif changed into another motif that this can be recognized by kinase B. So, these are the most simplest examples of these kinds of events that we are looking at.

(Refer Slide Time: 13:43)




There is also like further or like events that happen you know further part from the phosphorylation site. For example, if a mutation hits a kinase domain which contains the catalytic function of this kinase, it can also change the can lead to a to a aberrant kinase activity.

(Refer Slide Time: 14:05)

A selection of tools to study the impact of genomic mutations on signaling

Name	URL	Remarks	Ref
PhosphoPOINT	http://kinase.bioinformatics.tw (Link not working at the time of writing)	Human kinase interactome and phospho-protein database	2008
PhosphoVariant	Not available	Database for definite and possible variants changing phosphosites	2009
PhosSNP	http://phosnp.biocuckoo.org	Database of mutations predicted to impact phosphorylation status of proteins	2010
ActiveDriver	http://individual.utoronto.ca/jzhang/ActiveDriver/	Prediction of 'active' phosphosites in proteins that are specifically and significantly mutated in cancer genomes	2013
PTMvar	http://phosphosite.org	Database intersecting non-synonymous SNPs and PTM sites	2015
MIMP	http://mim.baderlab.org	Characterization of genetic variants that specifically alter kinase-binding sites in proteins	2015
ReKINect	http://rekinect.science.hmc.edu	Prediction of network attacking mutations (NAMs) from NGS data	2015
g2pDB	http://g2pdb.org	Database of auto-curated PTM sites mapped to their genomic locations	2016
ActiveDriverDB	http://www.activedriverdb.org	Database that annotates disease mutations and population variants through the lens of PTMs.	2018



So, here on this slide I have just presented a couple of tools. So, there is many tools out there already and you know people have started looking into that 10 years back already, but also recently there are lots of new developments. So, one so, it is actually the two tools we are we are going to have a closer look at. So, one tool is called MIMP and the other is called ActiveDriverDB, just want to highlight. So, this other tool here g2p database, genome to phosphoryl database, which has been developed in David Fenyo's lab as well.

Student: So, motif-X I tried to use that one tool. So, I like encountered a problem here. I have a phospho site and the residue phosphorylated like the serine is phosphohorylated at position 32 and I have the sequence, peptide sequence.

Yeah.

Student: but, the there it was asking which site is the phosphorylated like in the peptide might be that the second position is the fourth position.

Yes.

Student: but, in my sequence it was serine 32. So, I will now able to use the motif-X because I am biologist, I am not a programmer.

Yes.

Student: So, is any one of these tool, as you know all of these tools, is anyone of this tools solve my problem?

No, well. These tools do not solve your problem yeah. So, this is one step actually upstream of this type of analysis. So, many of these tools that actually take the raw mass spectrometry data and so, do the database search and create result reports on a phosphosite level let us say. So, many of these tools already have these kind of sequence windows in their result table. So, this is actually not a peptide sequence right. It is the sequence window which is always the same length.

So, in this case it is always 15 amino acids. It is the same this and the modification sizes in the center. So, this is something that many tools create like MaxQuant does it, like Spectrum mill does it I am not sure whether proteome discoverer does it as well.

Student: No, from proteome discoverer I got a peptide sequence.

Yes.

Student: And, with this serine 32.

Yes.

Student: and, might be my serine is at the fourth position.

Yes.

Student: So, I was not able to use the.

Yes.

Student: MIMP tool and this motif-X.


I see your problem. So, what you would have to do and you need to a hire a programmer that just takes the data and takes the database and creates the sequence window. So, or you just use another software. So, I can highly recommend MaxQuant ok.

(Refer Slide Time: 16:42)

Mutation impact on phosphorylation (MIMP)

- Predicts the impact of non-synonymous single-nucleotide variants (nsSNVs) on kinase-substrate interactions
 - Prediction of kinase binding affinities
 - Prediction of rewiring effects of nsSNVs
- Compares nsSNV effect in mutated (mut) and wild type (wt) sample
- Bayesian approach to construct position weight matrix (PWM) models of amino acid specificities of kinases

Wagih *et al.* Nature Methods. 2015 <http://mimp.baderlab.org/>



So, MIMP does exactly what we have just talked about. So, it predicts the impact of non-synonymous as on this as on kind of substrate interactions. So, it predicts kinase binding affinities and how or whether mutation re-wires protein or like phospho signaling networks.


And, it compares, so, that is basically like the two principal here. So, it compares the effect of mutated and wild type samples and one in a specific property of this tool is that it uses a patient approach to construct like these in this case position weight matrices which is probably which is very similar to these amino acid frequency matrix that we just looked at.

(Refer Slide Time: 17:31)

MIMP – key features

- 1) Kinase-binding models using known substrate sites
 - Software included pretrained models
- 2) Calculation of kinase binding score for a given phospho sequence
 - User-defined phosphosequences
 - PhosphoSitePlus
- 3) Mapping of cancer mutations to phosphosites
 - User-defined mutations
 - TCGA
- 4) Prediction of network rewiring events

Wagih *et al.* Nature Methods. 2015 <http://mimp.baderlab.org/>



So, this has been published in Nature Methods like 3 years back and there is an online tool which you can just use, but there is also like an R implementation of that package and this is actually what we are trying to use in the hands on sessions. So, I hope that we will be able to do that. So, we will see.

So, the key features is it first builds kinase binding models using known substrate sites. So, very highly curated very well known phosphorylation sites that have been determined to be a substrate of this particular kinase and use this information to build binding models. And, these models are already included in the software all right. So, there is nothing that you have to worry about.


So, then calculates for given phospho sequence that might be now coming you from your data set. It calculates the kinase binding score for each kinase. It calculates its score how likely it is that this force beside has been phosphorylated by the specific, but this specific kinase. And, again so, you can upload your own phospho sequences or you can just query all phospho site sequences and phospho site plus. Thus everybody know about this resource here for phosphoSite plus. So, I can again high recommend to check that out.

So, at the end of my slides there are like all references that I am going through here all included. So, you can just go through the papers and check them out yes. So, the first part is calculating kinase binding specificities and then it pulls in mutation data which again you can upload your own mutations or take or you can specifically look for TCGA mutations and then it does its prediction.

(Refer Slide Time: 19:16)

Points to Ponder

- SNPs are known to be the cause for many diseases but the molecular mechanisms involved are not known yet.
- Direct hit and proximal hit mutations can affect phosphosites leading to a loss of phosphorylation ability.
- Mutations in kinase domain can lead to aberrant activity of the kinase.



NPTEL IIT Bombay

So, today, in conclusion I hope you have learned that kinase activity gets affected due to the amino acids, which are present in the surrounding of the phosphorylation site. Hence, if we know the correlation of amino acid to phosphorylation is specificity then we may change the expression pattern of a gene we also heard that motif-X follows an iterative workflow which provides us reliable and confident amino acid sequence which could result for the phosphorylation regulation.

Mutations can lead to genesis modification or destruction of P-sites resulting in altered pathways. Dr. Krug also enlisted many tools which can be used for correlation between genomic mutations and signaling pathways. The next lecture will be the continuation of mutation and signaling by Dr. Karsten Krug.

Thank you.