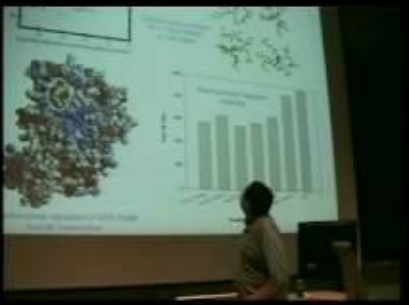
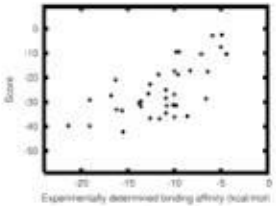
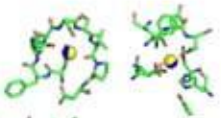


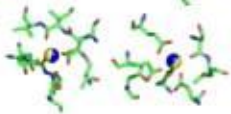
FUNCTION

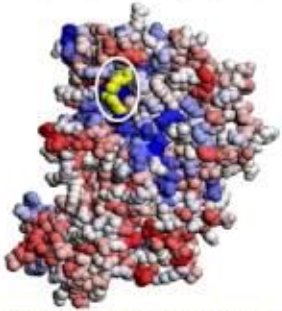


Ion binding energy prediction with a correlation of 0.7

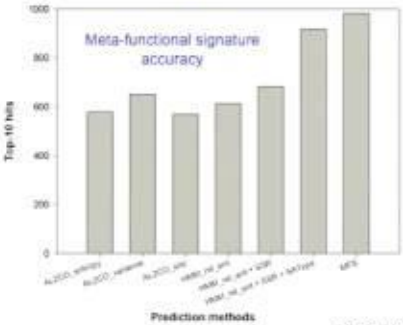


Calcium ions predicted to < 0.05 Å RMSD in 130 cases





Meta-functional signature for DXS model from *M. tuberculosis*



Wang/Cheng

1
00:00:06,280 --> 00:00:03,050
microbiology at the University of

2
00:00:08,810 --> 00:00:06,290
Washington is interested in becoming

3
00:00:12,770 --> 00:00:08,820
affiliated with the astrobiology program

4
00:00:15,919 --> 00:00:12,780
and he's the advisor of PhD advisor of

5
00:00:23,150 --> 00:00:15,929
Aaron Goldman and he'll be talking today

6
00:00:36,500 --> 00:00:23,160
on modeling proteomes so I decided to

7
00:00:38,389 --> 00:00:36,510
introduce can we get some later so the

8
00:00:39,979 --> 00:00:38,399
fundamental question that me and my

9
00:00:42,319 --> 00:00:39,989
group on Hansen's have been five years

10
00:00:44,420 --> 00:00:42,329
old trying to understand is how does

11
00:00:49,760 --> 00:00:44,430
Jesus organisms phosphites believer

12
00:00:52,069 --> 00:00:49,770
encourages character six I am and you

13
00:00:54,069 --> 00:00:52,079

know I've been told that this experience

14
00:00:56,990 --> 00:00:54,079
of astronomers and biologists and

15
00:00:59,389 --> 00:00:57,000
fundamentally believe that evolution

16
00:01:02,779 --> 00:00:59,399
understanding evolution is a way to

17
00:01:06,080 --> 00:01:02,789
understand how life occurred on earth

18
00:01:08,500 --> 00:01:06,090
and also how to design or see other

19
00:01:10,820 --> 00:01:08,510
planets but like that's it that's

20
00:01:13,310 --> 00:01:10,830
relevant to astrobiology and I think it

21
00:01:14,960 --> 00:01:13,320
is so actually so the title of my talk

22
00:01:16,820 --> 00:01:14,970
is king is a little bit because it's

23
00:01:18,530 --> 00:01:16,830
born with podium because I went back to

24
00:01:20,630 --> 00:01:18,540
some of them all the slides among older

25
00:01:24,859 --> 00:01:20,640
work because that i thought was more

26
00:01:27,980 --> 00:01:24,869
element and a focus mostly has been on

27
00:01:29,480 --> 00:01:27,990
proteins and probes I mean quickly under

28
00:01:32,990 --> 00:01:29,490
sculpture and putting I'll define all

29
00:01:36,710 --> 00:01:33,000
these terms but so so our focus has been

30
00:01:39,170 --> 00:01:36,720
on proteins and so i wanted i wanted i

31
00:01:42,380 --> 00:01:39,180
wanted to talk about that as i go

32
00:01:44,270 --> 00:01:42,390
through the talk I you know I was going

33
00:01:45,830 --> 00:01:44,280
to it a couple days ago and I was

34
00:01:48,139 --> 00:01:45,840
thinking maybe I should annotate every

35
00:01:49,940 --> 00:01:48,149
asked belgica aspect of it and then I

36
00:01:53,120 --> 00:01:49,950
thought hopefully I will be able to do

37
00:01:56,090 --> 00:01:53,130
it verbally I'll say me time I'd just

38
00:01:57,200 --> 00:01:56,100

been busy so let's start with the thing

39
00:01:58,340 --> 00:01:57,210
let's start with the fundamental

40
00:02:00,249 --> 00:01:58,350
question how does the genome of an

41
00:02:03,050 --> 00:02:00,259
argument specifies behavior capture

42
00:02:05,770 --> 00:02:03,060
characteristics if we can do that if we

43
00:02:07,999 --> 00:02:05,780
can understand that here on earth in

44
00:02:10,070 --> 00:02:08,009
extreme environments we can do that

45
00:02:10,779 --> 00:02:10,080
anywhere we can do it on Jupiter they

46
00:02:16,890 --> 00:02:10,789
can do that

47
00:02:19,479 --> 00:02:16,900
are speaking do it all no neptr so and

48
00:02:21,789 --> 00:02:19,489
the way we propose to answer this

49
00:02:24,789 --> 00:02:21,799
question is by doing something called

50
00:02:28,270 --> 00:02:24,799
modeling pull-ups the modeling part

51
00:02:30,420 --> 00:02:28,280
should be hopefully obvious modeling

52
00:02:32,410 --> 00:02:30,430
means that i do only computation work

53
00:02:34,839 --> 00:02:32,420
everything that we do this computer

54
00:02:37,179 --> 00:02:34,849
simulation you do not do any extra motor

55
00:02:39,580 --> 00:02:37,189
work but we do collaborate with extra

56
00:02:41,949 --> 00:02:39,590
Atlas who verifies our its module

57
00:02:44,729 --> 00:02:41,959
techniques i will give you a lot of data

58
00:02:47,530 --> 00:02:44,739
on that and that data may not be so

59
00:02:48,789 --> 00:02:47,540
immediately relevant to astrobiology but

60
00:02:52,599 --> 00:02:48,799
i want to make it very clear how

61
00:02:54,190 --> 00:02:52,609
relevant it would be to astrobiology

62
00:02:56,679 --> 00:02:54,200
sounds like what we are doing is

63
00:03:00,670 --> 00:02:56,689

developing a sense of self tools and

64
00:03:02,979 --> 00:03:00,680
techniques that is relevant to a huge

65
00:03:05,860 --> 00:03:02,989
number of disciplines and it's very

66
00:03:17,800 --> 00:03:05,870
broad in vera gel and what I mean by pro

67
00:03:20,140 --> 00:03:17,810
do so proteome in general to me is just

68
00:03:21,520 --> 00:03:20,150
a collection of proteins and then when

69
00:03:23,110 --> 00:03:21,530
you talk about interactome which is

70
00:03:25,360 --> 00:03:23,120
original title of my talk and you know

71
00:03:29,229 --> 00:03:25,370
to get into that it's the collection of

72
00:03:31,750 --> 00:03:29,239
the biologically relevant molecules if

73
00:03:34,809 --> 00:03:31,760
you want to call it that but in life in

74
00:03:37,149 --> 00:03:34,819
this planet almost everything functions

75
00:03:38,830 --> 00:03:37,159
our life are carried out the proteins so

76
00:03:41,170 --> 00:03:38,840
in this case what I've done is a

77
00:03:43,750 --> 00:03:41,180
proteome is a system and the system is

78
00:03:46,180 --> 00:03:43,760
circled by this blue box here and the

79
00:03:49,809 --> 00:03:46,190
proteins by themselves the individual

80
00:03:51,939 --> 00:03:49,819
objects are colored in black boxes so we

81
00:03:53,890 --> 00:03:51,949
don't know what they do so the first job

82
00:03:57,219 --> 00:03:53,900
is to understand what what each protein

83
00:03:58,960 --> 00:03:57,229
does on a molecular level and what I

84
00:04:01,210 --> 00:03:58,970
want to say is to give you an idea of

85
00:04:03,460 --> 00:04:01,220
what happens you know if you want to you

86
00:04:07,929 --> 00:04:03,470
all the complex organism in some other

87
00:04:09,759 --> 00:04:07,939
organic in some other environment let's

88
00:04:12,099 --> 00:04:09,769

say in extreme environments on earth or

89
00:04:14,170 --> 00:04:12,109
an extreme environment on the planet you

90
00:04:16,360 --> 00:04:14,180
really want to look at what what it

91
00:04:19,870 --> 00:04:16,370
takes there are about 60,000 open

92
00:04:24,250 --> 00:04:19,880
reading frames or yeah or proteins in

93
00:04:28,510 --> 00:04:26,620
rice is another bottle are you looking

94
00:04:31,030 --> 00:04:28,520
at very seriously one of my sources of

95
00:04:33,640 --> 00:04:31,040
funding said it's got 60,000 rice is one

96
00:04:37,870 --> 00:04:33,650
of the smaller applying genomes so

97
00:04:40,450 --> 00:04:37,880
things like onion and wheat and maize

98
00:04:45,070 --> 00:04:40,460
and so on have much much much much

99
00:04:46,480 --> 00:04:45,080
larger number of proteins so India

100
00:04:50,200 --> 00:04:46,490
understand why that is happening and

101
00:04:52,380 --> 00:04:50,210
then but but let's say we want to grow

102
00:04:55,120 --> 00:04:52,390
back to you know Jupiter there's about

103
00:04:56,860 --> 00:04:55,130
4,500 genes and there's something called

104
00:05:00,070 --> 00:04:56,870
a minimal Genome Project which i won't

105
00:05:01,660 --> 00:05:00,080
go into detail but that is trying to

106
00:05:03,760 --> 00:05:01,670
find the minimum set of proteins that

107
00:05:07,180 --> 00:05:03,770
electron argues arrived under specific

108
00:05:10,810 --> 00:05:07,190
conditions it could be a highly tomalak

109
00:05:13,480 --> 00:05:10,820
condition could be a highly variant

110
00:05:15,010 --> 00:05:13,490
condition but that that's that was

111
00:05:19,230 --> 00:05:15,020
actually initiated by craig Venter a

112
00:05:21,760 --> 00:05:19,240
tiger and that turns out to be about I

113
00:05:23,650 --> 00:05:21,770

think the estimate is turned out to be

114
00:05:24,760 --> 00:05:23,660
able to 400 to find genes or we were

115
00:05:26,770 --> 00:05:24,770
involved in trying to predict the

116
00:05:29,020 --> 00:05:26,780
functions and trying to get the proteome

117
00:05:31,120 --> 00:05:29,030
of that 400 genes and trying to put all

118
00:05:35,920 --> 00:05:31,130
together and I'll go into that into more

119
00:05:38,800 --> 00:05:35,930
detail so no matter what we do why am I

120
00:05:41,080 --> 00:05:38,810
saying it's 400 why do we even though

121
00:05:43,000 --> 00:05:41,090
there are 60,000 you and 6,000 in rice

122
00:05:47,770 --> 00:05:43,010
and even for thousands fall back here

123
00:05:50,590 --> 00:05:47,780
there are 400 because all of these

124
00:05:53,110 --> 00:05:50,600
proteins that are there together can be

125
00:05:54,400 --> 00:05:53,120
grouped into several thousand distinct

126
00:05:56,530 --> 00:05:54,410
signals families and this is where

127
00:05:58,330 --> 00:05:56,540
evolution comes in to cook so everything

128
00:06:03,900 --> 00:05:58,340
award from something else something else

129
00:06:07,270 --> 00:06:03,910
and so you have evolutionary divergence

130
00:06:10,510 --> 00:06:07,280
which is you know something involved and

131
00:06:12,220 --> 00:06:10,520
then you have some variants of that and

132
00:06:13,660 --> 00:06:12,230
that perform maybe different function

133
00:06:16,720 --> 00:06:13,670
may be different structures and so on

134
00:06:19,720 --> 00:06:16,730
and yellow Lucia convergence things that

135
00:06:21,370 --> 00:06:19,730
need to perform the same function and by

136
00:06:24,040 --> 00:06:21,380
chance they'll arrive with the same

137
00:06:28,630 --> 00:06:24,050
answer but that is by chance at least by

138
00:06:30,190 --> 00:06:28,640

current Darwinian ka so what the basic

139
00:06:31,990 --> 00:06:30,200
point here is that even though we have a

140
00:06:34,330 --> 00:06:32,000
large number of proteins and all these

141
00:06:36,220 --> 00:06:34,340
organisms these proteins can be grouped

142
00:06:36,770 --> 00:06:36,230
into several thousand distinct sequence

143
00:06:40,720 --> 00:06:36,780
family

144
00:06:44,270 --> 00:06:40,730
is that going in case till to my chair

145
00:06:47,360 --> 00:06:44,280
attractive I'm just toast wondering what

146
00:06:49,480 --> 00:06:47,370
a sequence family is just so there it is

147
00:06:51,260 --> 00:06:49,490
yeah similarities of sequences

148
00:06:54,379 --> 00:06:51,270
evolutionary relatedness I'm

149
00:06:57,379 --> 00:06:54,389
mathematical use yes yes correlation

150
00:06:59,810 --> 00:06:57,389
between it'sit's homology action the

151
00:07:04,010 --> 00:06:59,820
exact where is it homology yeah yeah so

152
00:07:06,590 --> 00:07:04,020
it's that well semaj your analogy since

153
00:07:09,620 --> 00:07:06,600
that they are either divergent or the

154
00:07:11,500 --> 00:07:09,630
conversion but this is more similar and

155
00:07:14,870 --> 00:07:11,510
they're performing they look similar

156
00:07:18,409 --> 00:07:14,880
sequences looks able at the sequence

157
00:07:21,190 --> 00:07:18,419
level again I'll go a little bit into

158
00:07:23,150 --> 00:07:21,200
what proteins really are in a minute but

159
00:07:24,800 --> 00:07:23,160
the first thing if you want to

160
00:07:26,570 --> 00:07:24,810
understand you want to remove those

161
00:07:28,370 --> 00:07:26,580
black boxes are those are black boxes

162
00:07:30,350 --> 00:07:28,380
for a reason because we don't understand

163
00:07:32,030 --> 00:07:30,360

what they do so we want to understand

164
00:07:34,820 --> 00:07:32,040
what it look like because everything

165
00:07:36,469 --> 00:07:34,830
that happens in life and physics I mean

166
00:07:39,469 --> 00:07:36,479
it's all atomic interactions so if you

167
00:07:41,360 --> 00:07:39,479
think about as is the universe we're

168
00:07:43,040 --> 00:07:41,370
just a bunch of atoms you know in

169
00:07:45,590 --> 00:07:43,050
collections together and we are in

170
00:07:47,780 --> 00:07:45,600
tracking in some ways and there right

171
00:07:50,659 --> 00:07:47,790
now ain't right mean in particularly and

172
00:07:52,279 --> 00:07:50,669
so in going back to these molecular

173
00:07:55,630 --> 00:07:52,289
systems that we are talking about our

174
00:07:58,040 --> 00:07:55,640
proteins we want to send what these

175
00:08:00,020 --> 00:07:58,050
proteins look like at our atomic level

176
00:08:02,750 --> 00:08:00,030
you want to send the precise coordinates

177
00:08:04,159 --> 00:08:02,760
you want understand every atom or at

178
00:08:06,230 --> 00:08:04,169
least at least the heavy atoms that's

179
00:08:08,930 --> 00:08:06,240
what we focus on and so we want to send

180
00:08:11,719 --> 00:08:08,940
the position of every atom and here's

181
00:08:13,719 --> 00:08:11,729
the interesting part about this even

182
00:08:16,100 --> 00:08:13,729
though there are many many thousands

183
00:08:18,050 --> 00:08:16,110
several thousand distinct signal

184
00:08:20,630 --> 00:08:18,060
families the number of structures

185
00:08:22,159 --> 00:08:20,640
structural families the structures that

186
00:08:24,230 --> 00:08:22,169
are similar to each other by some

187
00:08:27,950 --> 00:08:24,240
measure say root mean square deviation

188
00:08:31,310 --> 00:08:27,960

or do any superimpose them they are very

189
00:08:33,020 --> 00:08:31,320
few actual there was a paper in Nature

190
00:08:35,510 --> 00:08:33,030
proposed by size shirt here who said

191
00:08:38,120 --> 00:08:35,520
there's only a thousand but it's in the

192
00:08:41,420 --> 00:08:38,130
order of that much we we probably have

193
00:08:43,700 --> 00:08:41,430
about 600 unique protein families in

194
00:08:45,949 --> 00:08:43,710
terms of structure and every year is

195
00:08:46,750 --> 00:08:45,959
probably one or two being discovered so

196
00:08:48,460 --> 00:08:46,760
we are the tail

197
00:08:50,650 --> 00:08:48,470
of the distribution so we discovered

198
00:08:52,960 --> 00:08:50,660
pretty much all the sequence families

199
00:08:55,210 --> 00:08:52,970
that we can using standard experimental

200
00:08:57,280 --> 00:08:55,220
techniques that is X a diffraction a

201
00:08:58,510 --> 00:08:57,290
mass spectroscopy there might be a class

202
00:09:01,360 --> 00:08:58,520
of protein that we haven't discovered

203
00:09:03,700 --> 00:09:01,370
yet or structures yet using new methods

204
00:09:06,340 --> 00:09:03,710
but but for using current ex-model

205
00:09:08,560 --> 00:09:06,350
techniques there are only a few thousand

206
00:09:11,170 --> 00:09:08,570
listing structural force what that means

207
00:09:13,360 --> 00:09:11,180
is that evolution is reusing this these

208
00:09:16,870 --> 00:09:13,370
structures again and again to perform

209
00:09:18,280 --> 00:09:16,880
many many and get to the next slide to

210
00:09:24,430 --> 00:09:18,290
perform many many different functions

211
00:09:28,180 --> 00:09:24,440
and that's so that's in all species that

212
00:09:30,010 --> 00:09:28,190
use that weapon this is over this is

213
00:09:32,170 --> 00:09:30,020

over yeah every single Freddie I mean

214
00:09:34,720 --> 00:09:32,180
millions and millions of sequences every

215
00:09:36,760 --> 00:09:34,730
every protein a human may say every

216
00:09:39,010 --> 00:09:36,770
protein honor that has been sequenced

217
00:09:42,400 --> 00:09:39,020
every gene in the protein others being

218
00:09:44,680 --> 00:09:42,410
safe ones that has been taught as it as

219
00:09:46,690 --> 00:09:44,690
a protein as an open reading frame so

220
00:09:50,410 --> 00:09:46,700
everything there's only a few thousand

221
00:09:52,810 --> 00:09:50,420
structural shapes that they are not so

222
00:09:54,100 --> 00:09:52,820
the structural atomic ships so that's

223
00:09:58,330 --> 00:09:54,110
where we are working on we are working

224
00:10:00,520 --> 00:09:58,340
at at alone at the atomic shape level so

225
00:10:03,640 --> 00:10:00,530
you might think that that would mean

226
00:10:05,200 --> 00:10:03,650
that would be to make up from ZZ but

227
00:10:07,630 --> 00:10:05,210
actually doesn't it actually makes up

228
00:10:09,790 --> 00:10:07,640
front very very hard because even though

229
00:10:11,860 --> 00:10:09,800
the sequence there only several thousand

230
00:10:13,600 --> 00:10:11,870
sequins families and a few thousand

231
00:10:15,340 --> 00:10:13,610
structural force maybe a thousand

232
00:10:16,720 --> 00:10:15,350
structural folds we are about 600 right

233
00:10:18,250 --> 00:10:16,730
now and there's a telling the

234
00:10:21,190 --> 00:10:18,260
dissipation or it might be a long tail I

235
00:10:23,620 --> 00:10:21,200
don't know but tens and thousands of

236
00:10:27,370 --> 00:10:23,630
millions of functions right now proteins

237
00:10:29,320 --> 00:10:27,380
too many many many many in HX let me

238
00:10:31,870 --> 00:10:29,330

talk right now and they do everything

239
00:10:36,310 --> 00:10:31,880
and even though they have the same shape

240
00:10:38,170 --> 00:10:36,320
they look the same bye-bye any if you

241
00:10:42,550 --> 00:10:38,180
look at that by they look to say to you

242
00:10:44,830 --> 00:10:42,560
but by by some mechanism which we can

243
00:10:47,650 --> 00:10:44,840
rationalize but if you look carefully

244
00:10:49,900 --> 00:10:47,660
enough they do different different

245
00:10:52,030 --> 00:10:49,910
things so again it goes back to the

246
00:10:55,450 --> 00:10:52,040
sport of evolution they're using the

247
00:10:56,759 --> 00:10:55,460
same sequence and same structure again

248
00:10:59,369 --> 00:10:56,769
and again to do

249
00:11:00,929 --> 00:10:59,379
many many different things and so it

250
00:11:02,429 --> 00:11:00,939
goes back to the issue of minimal genome

251
00:11:04,679 --> 00:11:02,439
Anasta value so if you want to see the

252
00:11:06,809 --> 00:11:04,689
planet over the life what you want to do

253
00:11:12,030 --> 00:11:06,819
is get the normal set of structures that

254
00:11:17,069 --> 00:11:12,040
you need for for an organism to function

255
00:11:19,619 --> 00:11:17,079
problem now so far our simplified the

256
00:11:22,769 --> 00:11:19,629
view a little bit i'm talking about

257
00:11:24,090 --> 00:11:22,779
individual proteins individual genes if

258
00:11:26,549 --> 00:11:24,100
you want to call it most familiar people

259
00:11:28,919 --> 00:11:26,559
of my jeans rejean usually transcribes

260
00:11:31,169 --> 00:11:28,929
are protein and we've talked about

261
00:11:34,289 --> 00:11:31,179
individual protein function and

262
00:11:36,319 --> 00:11:34,299
structure and sequence but what happens

263
00:11:38,850 --> 00:11:36,329

is it's it's an interconnected system

264
00:11:40,729 --> 00:11:38,860
it's a huge interconnected system and

265
00:11:43,189 --> 00:11:40,739
that's what we are trying to get at so

266
00:11:45,840 --> 00:11:43,199
what matters in this interconnected

267
00:11:49,049 --> 00:11:45,850
expression so we want to know how many

268
00:11:51,449 --> 00:11:49,059
copies of each protein or each

269
00:11:53,729 --> 00:11:51,459
functional unit there are you can think

270
00:11:55,530 --> 00:11:53,739
of H proteins of community we want to

271
00:11:57,179 --> 00:11:55,540
know how many copies of each functional

272
00:11:59,100 --> 00:11:57,189
unit are and there are different

273
00:12:01,019 --> 00:11:59,110
expression patterns based on time and

274
00:12:05,519 --> 00:12:01,029
location based on the development of the

275
00:12:06,960 --> 00:12:05,529
organism in inert at least and then we

276
00:12:09,239 --> 00:12:06,970
want to understand how they all interact

277
00:12:12,179 --> 00:12:09,249
with so many copies of all these things

278
00:12:14,160 --> 00:12:12,189
you want to understand exactly how all

279
00:12:16,079 --> 00:12:14,170
these copies interact with each other to

280
00:12:18,749 --> 00:12:16,089
perform what we call the organism

281
00:12:20,819 --> 00:12:18,759
performant back you know and that's what

282
00:12:22,410 --> 00:12:20,829
i mean by an how does the genome of our

283
00:12:25,259 --> 00:12:22,420
businesses fights behavior encourage

284
00:12:26,910 --> 00:12:25,269
text what i want to say too is very

285
00:12:29,400 --> 00:12:26,920
fundamental basic point is that the

286
00:12:31,980 --> 00:12:29,410
interaction and expression then were

287
00:12:33,809 --> 00:12:31,990
copies and how they interact are very

288
00:12:36,780 --> 00:12:33,819

interdependent with a molecular

289
00:12:39,780 --> 00:12:36,790
structure and function so you are trying

290
00:12:41,909 --> 00:12:39,790
to relate sis when mean in ab article

291
00:12:46,679 --> 00:12:41,919
sense we call this aspect of it systems

292
00:12:48,749 --> 00:12:46,689
biology and then from the other side we

293
00:12:50,549 --> 00:12:48,759
call biophysics biochemistry whatever

294
00:12:52,199 --> 00:12:50,559
but what we are trying to do is relate

295
00:12:54,780 --> 00:12:52,209
the two things together and this is what

296
00:12:56,850 --> 00:12:54,790
my group is trying to do and we have

297
00:13:01,309 --> 00:12:56,860
developed a set of technologies and

298
00:13:04,650 --> 00:13:01,319
tools that Adam is working partly on

299
00:13:06,720 --> 00:13:04,660
that that that tries to get at it and

300
00:13:08,639 --> 00:13:06,730
what I'm going to do now for the

301
00:13:09,530 --> 00:13:08,649
remainder of the talk is actually going

302
00:13:14,990 --> 00:13:09,540
to

303
00:13:16,670 --> 00:13:15,000
results rather than explaining what the

304
00:13:19,790 --> 00:13:16,680
tools to because that would that that

305
00:13:21,439 --> 00:13:19,800
itself would take a long time and then I

306
00:13:23,600 --> 00:13:21,449
want to talk about some applications of

307
00:13:26,689 --> 00:13:23,610
these tools and the application right

308
00:13:31,009 --> 00:13:26,699
now are not again astrobiological in any

309
00:13:33,620 --> 00:13:31,019
sense but I want to talk about how they

310
00:13:35,600 --> 00:13:33,630
can be asked about you and that's again

311
00:13:39,230 --> 00:13:35,610
something that that Adam has been

312
00:13:41,480 --> 00:13:39,240
challenged with okay so let's go back to

313
00:13:44,480 --> 00:13:41,490

the basics so if you're not following up

314
00:13:46,670 --> 00:13:44,490
to this point just kind of give you a

315
00:13:49,639 --> 00:13:46,680
very big background because we ate did

316
00:13:51,050 --> 00:13:49,649
ask me to and I normally said I don't

317
00:13:53,629 --> 00:13:51,060
like doing self my audiences with this

318
00:13:56,840 --> 00:13:53,639
but I want change this is that there is

319
00:14:00,110 --> 00:13:56,850
a gene and for when there are many genes

320
00:14:03,100 --> 00:14:00,120
in your in your in your body there are

321
00:14:05,840 --> 00:14:03,110
many genes that code for many proteins

322
00:14:07,910 --> 00:14:05,850
the coral system is pretty well worked

323
00:14:09,410 --> 00:14:07,920
out and one of the first projects my

324
00:14:11,389 --> 00:14:09,420
undergrad you did is actually trying to

325
00:14:13,400 --> 00:14:11,399
figure out why did the colon system come

326
00:14:16,129 --> 00:14:13,410
out the way it did there's no publisher

327
00:14:17,750 --> 00:14:16,139
I won't go into that into detail but it

328
00:14:20,720 --> 00:14:17,760
turns out that there are reasons for

329
00:14:24,309 --> 00:14:20,730
that but what happens is that three base

330
00:14:28,189 --> 00:14:24,319
pairs the jeans are composed of DNA and

331
00:14:30,500 --> 00:14:28,199
there are four types of DNA nucleotides

332
00:14:33,170 --> 00:14:30,510
on the oxygen of the deoxyribose

333
00:14:35,090 --> 00:14:33,180
nucleotides 80 c and g will call them

334
00:14:37,819 --> 00:14:35,100
that you can look at it as letters on

335
00:14:41,180 --> 00:14:37,829
string it doesn't matter and three pairs

336
00:14:44,360 --> 00:14:41,190
of those I mean pair decide three sets

337
00:14:47,030 --> 00:14:44,370
of those code for one amino acid in a

338
00:14:49,100 --> 00:14:47,040

protein and again we're focused on

339
00:14:51,610 --> 00:14:49,110
protein so we work with work at this

340
00:14:53,780 --> 00:14:51,620
level and there are 20 amino acids and

341
00:14:56,360 --> 00:14:53,790
they perform they have different

342
00:14:58,009 --> 00:14:56,370
chemical groups they perform different

343
00:15:00,439 --> 00:14:58,019
functions and that's what gives a

344
00:15:02,540 --> 00:15:00,449
protein that's what makes a protein 11

345
00:15:05,000 --> 00:15:02,550
protein different from other protein so

346
00:15:06,800 --> 00:15:05,010
if you have identical proteins that have

347
00:15:08,179 --> 00:15:06,810
all the minuses the same day we should

348
00:15:10,819 --> 00:15:08,189
perform the same function the same

349
00:15:12,860 --> 00:15:10,829
structure and look the same but if I you

350
00:15:14,329 --> 00:15:12,870
know a single change a single mutation

351
00:15:19,189 --> 00:15:14,339
it can cause the disease in your body

352
00:15:20,720 --> 00:15:19,199
that could kill you so so there there

353
00:15:22,460 --> 00:15:20,730
are 20 of mine acid and again

354
00:15:24,889 --> 00:15:22,470
20 amino acids are indicated by single

355
00:15:27,949 --> 00:15:24,899
letter codes like Ellis for leucine and

356
00:15:29,930 --> 00:15:27,959
case for lysine I won't go into detail

357
00:15:33,069 --> 00:15:29,940
on those but just assume that each each

358
00:15:36,889 --> 00:15:33,079
each amino acids encodes a different

359
00:15:39,829 --> 00:15:36,899
chemical function what happens and this

360
00:15:41,480 --> 00:15:39,839
is again a very simplified view of what

361
00:15:45,019 --> 00:15:41,490
we taught biology was at least what

362
00:15:46,790 --> 00:15:45,029
years ago is that this sequence genius

363
00:15:49,600 --> 00:15:46,800

transcribed and translated into this

364
00:15:52,550 --> 00:15:49,610
protein sequence and it produces a

365
00:15:53,990 --> 00:15:52,560
protein and when it releases proteins

366
00:15:55,670 --> 00:15:54,000
and none for lots it and this is what

367
00:15:59,480 --> 00:15:55,680
really an a-minus it kind of startled

368
00:16:01,910 --> 00:15:59,490
looks like say you had this a very basic

369
00:16:04,189 --> 00:16:01,920
structure to its got what we call a main

370
00:16:06,410 --> 00:16:04,199
chain it's a linear chain and then

371
00:16:07,970 --> 00:16:06,420
studies side group to it that's what

372
00:16:09,980 --> 00:16:07,980
makes a massive difference so it's got

373
00:16:12,199 --> 00:16:09,990
this is a lysine right here this is a

374
00:16:13,790 --> 00:16:12,209
carboxyl right here and so on and that's

375
00:16:16,550 --> 00:16:13,800
what makes each owner that's different

376
00:16:20,180 --> 00:16:16,560
and we assume that when the protein is

377
00:16:22,699 --> 00:16:20,190
made it looks like this and we all also

378
00:16:23,960 --> 00:16:22,709
know that that this is the case enfant

379
00:16:26,180 --> 00:16:23,970
someone's nobel prize for showing this

380
00:16:28,160 --> 00:16:26,190
is the case but in general protein has

381
00:16:30,230 --> 00:16:28,170
to be folded and unfolded and refolded

382
00:16:32,269 --> 00:16:30,240
and fold it again to be transported

383
00:16:34,579 --> 00:16:32,279
other places and so on so this happens a

384
00:16:37,759 --> 00:16:34,589
lot so when you finish a maid or maybe

385
00:16:43,579 --> 00:16:37,769
doing certain conditions unfolded what

386
00:16:44,660 --> 00:16:43,589
happens we also know again well what I

387
00:16:46,670 --> 00:16:44,670
want to say something will the

388
00:16:48,920 --> 00:16:46,680

characters of the characteristics of

389
00:16:50,870 --> 00:16:48,930
this state is that it's not unique so

390
00:16:53,059 --> 00:16:50,880
this is fluctuating between many many

391
00:16:54,470 --> 00:16:53,069
many many different shapes again we are

392
00:16:56,750 --> 00:16:54,480
talking about our Tomic love the shapes

393
00:16:58,790 --> 00:16:56,760
and I'll illustrate that Dominic this is

394
00:17:00,319 --> 00:16:58,800
highly mobile and it's inactive so this

395
00:17:02,329 --> 00:17:00,329
is not a functional form of this protein

396
00:17:04,610 --> 00:17:02,339
this doesn't do anything and it's

397
00:17:06,919 --> 00:17:04,620
expanded usually and it's very good

398
00:17:08,569 --> 00:17:06,929
there's no order to it there's nothing

399
00:17:12,230 --> 00:17:08,579
that you can look at this and say oh

400
00:17:15,710 --> 00:17:12,240
look this is what it does so what

401
00:17:17,659 --> 00:17:15,720
happens nature again a very very

402
00:17:20,840 --> 00:17:17,669
simplified view and this is the problem

403
00:17:23,299 --> 00:17:20,850
that you're working on is that this this

404
00:17:25,549 --> 00:17:23,309
chain of amino acids change there's

405
00:17:28,909 --> 00:17:25,559
actually a bunch of atoms any chat then

406
00:17:30,799 --> 00:17:28,919
I've shown you it's a stick figure so

407
00:17:32,240 --> 00:17:30,809
this is actually a ball right here is a

408
00:17:33,240 --> 00:17:32,250
ball right here is ball right here with

409
00:17:36,510 --> 00:17:33,250
the Radia of an actor

410
00:17:38,250 --> 00:17:36,520
and this this chain spontaneously

411
00:17:40,440 --> 00:17:38,260
self-organized the time together the

412
00:17:43,050 --> 00:17:40,450
second less than a second milliseconds

413
00:17:45,990 --> 00:17:43,060

even the most proteins that the me off

414
00:17:52,350 --> 00:17:46,000
at least again that are being experiment

415
00:17:55,050 --> 00:17:52,360
like a price into this need to

416
00:17:59,400 --> 00:17:55,060
biologically or lemon state I mean what

417
00:18:02,340 --> 00:17:59,410
do I mean by that sorry yeah so into

418
00:18:05,070 --> 00:18:02,350
this native rg11 state so you have the

419
00:18:07,320 --> 00:18:05,080
carbon atoms are colored them in gray

420
00:18:10,950 --> 00:18:07,330
right here the oxygen atoms are colored

421
00:18:13,230 --> 00:18:10,960
in red and the nitrogen atoms apart so

422
00:18:15,390 --> 00:18:13,240
this is what happens within within a

423
00:18:17,940 --> 00:18:15,400
second right when the protein is named

424
00:18:19,800 --> 00:18:17,950
and you let it fold up and this is what

425
00:18:21,690 --> 00:18:19,810
happens why are you very quickly that's

426
00:18:24,720 --> 00:18:21,700
an evolutionary process that has been

427
00:18:27,690 --> 00:18:24,730
optimized by for most proteins for

428
00:18:29,520 --> 00:18:27,700
billions of years of evolution I mean

429
00:18:30,980 --> 00:18:29,530
that's that's that's and that's what

430
00:18:33,810 --> 00:18:30,990
everything is derived from that's why

431
00:18:36,660 --> 00:18:33,820
structure is doing this is a very very

432
00:18:41,220 --> 00:18:36,670
hard problem for any organism let all of

433
00:18:44,670 --> 00:18:41,230
us and so so this is why structure is so

434
00:18:46,200 --> 00:18:44,680
conserved among most organisms but

435
00:18:47,460 --> 00:18:46,210
looking at the protein like this with

436
00:18:49,410 --> 00:18:47,470
the ball of acting like that doesn't

437
00:18:51,780 --> 00:18:49,420
tell you anything I mean in the sense I

438
00:18:53,870 --> 00:18:51,790

doesn't reveal what what the proteins

439
00:18:56,910 --> 00:18:53,880
about so we look at in a very abstract

440
00:18:59,670 --> 00:18:56,920
view and abstract view is connecting the

441
00:19:02,100 --> 00:18:59,680
C alpha atoms or the carbon main chain

442
00:19:05,280 --> 00:19:02,110
atoms on Adam that's common among all

443
00:19:06,780 --> 00:19:05,290
the all the these are minor acids and we

444
00:19:08,250 --> 00:19:06,790
connect them together and we also color

445
00:19:11,250 --> 00:19:08,260
the direction of change from blue to red

446
00:19:12,750 --> 00:19:11,260
so we okay look at it like this then you

447
00:19:15,570 --> 00:19:12,760
can start seeing several features of

448
00:19:17,220 --> 00:19:15,580
this body what are they it's a very

449
00:19:19,200 --> 00:19:17,230
unique shape this is a transcription

450
00:19:21,810 --> 00:19:19,210
factor so it's a real protein I'm not

451
00:19:23,820 --> 00:19:21,820
making it up so this is a transcription

452
00:19:26,880 --> 00:19:23,830
factor of transcribed other protein mix

453
00:19:30,240 --> 00:19:26,890
of the proteins thousand a caliper so it

454
00:19:31,860 --> 00:19:30,250
has a very unique shape if you unfold it

455
00:19:34,620 --> 00:19:31,870
and refold it fall back into the same

456
00:19:36,570 --> 00:19:34,630
state so again like I said intensive on

457
00:19:39,330 --> 00:19:36,580
an all purpose for this it's very

458
00:19:40,830 --> 00:19:39,340
precisely or it it's stable and its

459
00:19:43,230 --> 00:19:40,840
function this is a functional formula

460
00:19:45,690 --> 00:19:43,240
protein and as you can see is cloud

461
00:19:48,509 --> 00:19:45,700
layer and compact like unlike the

462
00:19:50,580 --> 00:19:48,519
unfolded slow and it's got these regular

463
00:19:52,649 --> 00:19:50,590

substructures what we call he sees and

464
00:19:55,830 --> 00:19:52,659
sheets so right here you're looking at

465
00:19:57,149 --> 00:19:55,840
helix head down I'm here looking at

466
00:20:01,379 --> 00:19:57,159
helix right here and you're looking at a

467
00:20:03,720 --> 00:20:01,389
helix right there and then we call these

468
00:20:05,789 --> 00:20:03,730
things strengths and you know that

469
00:20:08,190 --> 00:20:05,799
there's there's some semantic like when

470
00:20:10,529 --> 00:20:08,200
we take that but we call them strands

471
00:20:12,389 --> 00:20:10,539
and these trends archaea hydrogen bond

472
00:20:14,700 --> 00:20:12,399
together and they form what we call a

473
00:20:16,769 --> 00:20:14,710
beta sheet so it is alpha hitter sees

474
00:20:21,919 --> 00:20:16,779
and their sheets and use that knowledge

475
00:20:24,960 --> 00:20:21,929
in part two to do all our predictions so

476
00:20:27,149 --> 00:20:24,970
surrender the rare the blue ends is just

477
00:20:31,019 --> 00:20:27,159
to guide the I within the spectrum of

478
00:20:33,779 --> 00:20:31,029
colors as to the order of things exactly

479
00:20:36,450 --> 00:20:33,789
so that there is an order to the protein

480
00:20:39,180 --> 00:20:36,460
right and it goes from n to the C terms

481
00:20:42,060 --> 00:20:39,190
when the protein synthesize you it is

482
00:20:44,639 --> 00:20:42,070
synthesized like that a decent size is

483
00:20:46,529 --> 00:20:44,649
that order so you start with blue we

484
00:20:48,810 --> 00:20:46,539
started blue the first blow right there

485
00:20:50,730 --> 00:20:48,820
and then the next one will added and

486
00:20:53,669 --> 00:20:50,740
next one is added next one at it and

487
00:20:55,379 --> 00:20:53,679
that's done by the proteins so and

488
00:20:57,600 --> 00:20:55,389

actually mess in your Ernie and so on

489
00:21:00,299 --> 00:20:57,610
it's a complex process but that's that's

490
00:21:03,240 --> 00:21:00,309
part of the ribosome but but yeah so

491
00:21:05,700 --> 00:21:03,250
that's exactly right so the the chain

492
00:21:08,820 --> 00:21:05,710
goes from aim to see what we call

493
00:21:11,779 --> 00:21:08,830
internal to the c-terminal so the ideal

494
00:21:13,860 --> 00:21:11,789
the the bonds are added in one direction

495
00:21:16,200 --> 00:21:13,870
they do not go in the reverse direction

496
00:21:17,970 --> 00:21:16,210
as far as I know and that will be a min

497
00:21:21,389 --> 00:21:17,980
an amazing discovery somebody found out

498
00:21:23,399 --> 00:21:21,399
that a 20 of the rails Jewess direction

499
00:21:27,930 --> 00:21:23,409
but back to last robotic irrelevance

500
00:21:29,789 --> 00:21:27,940
vehicle engineer we we as humans we can

501
00:21:31,200 --> 00:21:29,799
do a lot of things but we can engineer

502
00:21:33,120 --> 00:21:31,210
that we can make things go in the

503
00:21:36,930 --> 00:21:33,130
reverse direction but the way nature

504
00:21:38,700 --> 00:21:36,940
have selected things as Lucien alee it

505
00:21:41,009 --> 00:21:38,710
seems to be evolutionary more efficient

506
00:21:43,500 --> 00:21:41,019
all right has chosen randomly to pick

507
00:21:47,070 --> 00:21:43,510
one way and that is n to see it goes in

508
00:21:49,259 --> 00:21:47,080
one direction and with all this I boy is

509
00:21:54,960 --> 00:21:49,269
available over if you unfold it and

510
00:21:57,600 --> 00:21:54,970
refolded it you know that's I'm

511
00:21:58,890 --> 00:21:57,610
simplifying the picture a lot here but

512
00:22:01,890 --> 00:21:58,900
yes it would

513
00:22:03,480 --> 00:22:01,900

you said I mean this is if you you know

514
00:22:06,090 --> 00:22:03,490
somebody want a Nobel Prize for that

515
00:22:08,160 --> 00:22:06,100
showing that for showing exactly that

516
00:22:11,220 --> 00:22:08,170
I'm Turing exactly that question that

517
00:22:13,350 --> 00:22:11,230
you unfold Barney's and and it's an

518
00:22:15,780 --> 00:22:13,360
enzyme and it tree falls back to the

519
00:22:19,440 --> 00:22:15,790
same state unfolded it falls back and

520
00:22:22,380 --> 00:22:19,450
fold it refers back so yeah I need to

521
00:22:25,200 --> 00:22:22,390
fold in that direction but it's not

522
00:22:26,940 --> 00:22:25,210
clear that the well-done proteins made

523
00:22:29,310 --> 00:22:26,950
in addiction but how it actually folds

524
00:22:31,050 --> 00:22:29,320
actually not so clear so so that's

525
00:22:33,780 --> 00:22:31,060
actually that's a very tricky question

526
00:22:36,930 --> 00:22:33,790
because you're saying how it's made

527
00:22:40,800 --> 00:22:36,940
versus how its folding and that's that's

528
00:22:43,070 --> 00:22:40,810
that's a distinction that I don't think

529
00:22:45,210 --> 00:22:43,080
anyone knows the answer to that's

530
00:22:47,640 --> 00:22:45,220
probably one of the most fundamental and

531
00:22:49,560 --> 00:22:47,650
solve problems and biology how does it

532
00:22:52,080 --> 00:22:49,570
how does a protein fold and we are

533
00:22:54,510 --> 00:22:52,090
trying to answer that right so but but

534
00:22:56,340 --> 00:22:54,520
the way it's made is definitely in to

535
00:22:59,100 --> 00:22:56,350
see every protein this in the scene

536
00:23:00,920 --> 00:22:59,110
words that mean off is made anti-seize

537
00:23:03,570 --> 00:23:00,930
never remain in the reverse direction

538
00:23:05,640 --> 00:23:03,580

okay and there's never like you know

539
00:23:07,740 --> 00:23:05,650
parts all protein made here and positive

540
00:23:10,050 --> 00:23:07,750
14 made here and they join together and

541
00:23:11,640 --> 00:23:10,060
form a single chain that but also you

542
00:23:13,950 --> 00:23:11,650
never know I mean they might form dimers

543
00:23:16,410 --> 00:23:13,960
but there are single chains one single

544
00:23:18,720 --> 00:23:16,420
long chain and they're all connected by

545
00:23:24,840 --> 00:23:18,730
covalent bonds that's that's what I mean

546
00:23:26,700 --> 00:23:24,850
by that well by a single chain so that

547
00:23:28,800 --> 00:23:26,710
is the that is a problem that we're

548
00:23:31,200 --> 00:23:28,810
dealing it so woody woody product

549
00:23:34,890 --> 00:23:31,210
prescription can we take that sickness

550
00:23:37,320 --> 00:23:34,900
and can we predict the structure and

551
00:23:39,540 --> 00:23:37,330
that's why I spent a lot of my life

552
00:23:44,730 --> 00:23:39,550
working on more than half my life I

553
00:23:47,010 --> 00:23:44,740
would say at this point yeah and it's

554
00:23:48,420 --> 00:23:47,020
it's something that people spend more

555
00:23:51,900 --> 00:23:48,430
than 50 years on it's an unsolved

556
00:23:53,700 --> 00:23:51,910
problem but we are making progress and

557
00:23:56,400 --> 00:23:53,710
we're very good at it and we're getting

558
00:23:58,020 --> 00:23:56,410
very very not just my group I mean David

559
00:24:01,920 --> 00:23:58,030
Baker is another profits in biochemistry

560
00:24:04,230 --> 00:24:01,930
the top leaders world in doing this kind

561
00:24:05,670 --> 00:24:04,240
of thing and in the Pacific Northwest we

562
00:24:07,740 --> 00:24:05,680
are probably the only two people who can

563
00:24:09,480 --> 00:24:07,750

do this and in the world that's probably

564
00:24:10,990 --> 00:24:09,490
about a handful of people about five to

565
00:24:15,310 --> 00:24:11,000
ten people who can do this

566
00:24:16,780 --> 00:24:15,320
so how can we measure what we're doing

567
00:24:19,300 --> 00:24:16,790
is right there's the Rex mental

568
00:24:20,920 --> 00:24:19,310
techniques x-ray diffraction and mass

569
00:24:23,500 --> 00:24:20,930
spectroscopy that tell you what a

570
00:24:25,840 --> 00:24:23,510
protein looks like and I will not go

571
00:24:27,640 --> 00:24:25,850
into details on that but but distrust me

572
00:24:30,160 --> 00:24:27,650
on that when I say that there are ways

573
00:24:33,580 --> 00:24:30,170
to look at the protein structure atomic

574
00:24:38,980 --> 00:24:33,590
level detail and so that's a gold

575
00:24:41,440 --> 00:24:38,990
standard and then we we can do

576
00:24:45,970 --> 00:24:41,450
competition for small proteins I say

577
00:24:54,340 --> 00:24:45,980
small I mean 100 250 amino acids that is

578
00:24:56,290 --> 00:24:54,350
about 300g base pairs DNA bases we can

579
00:24:59,740 --> 00:24:56,300
actually predict the structure to p high

580
00:25:02,740 --> 00:24:59,750
accuracy in general on average so it go

581
00:25:04,960 --> 00:25:02,750
from 3 angstroms to about 6 ounces so

582
00:25:07,270 --> 00:25:04,970
when I say winning is a measuring stand

583
00:25:09,490 --> 00:25:07,280
people think of it as a resolution but

584
00:25:11,380 --> 00:25:09,500
it's actually accuracy of the gold

585
00:25:12,610 --> 00:25:11,390
standard the measure of the accuracy of

586
00:25:14,380 --> 00:25:12,620
the gold standard to what we are

587
00:25:16,330 --> 00:25:14,390
predicting it's a deviation from that

588
00:25:19,060 --> 00:25:16,340

it's a root mean squared deviation that

589
00:25:21,400 --> 00:25:19,070
mean usually the main chain atoms but

590
00:25:23,800 --> 00:25:21,410
but it can be anything can we all atoms

591
00:25:25,690 --> 00:25:23,810
but does this the spot will still stand

592
00:25:29,590 --> 00:25:25,700
let's so see how far must be in this

593
00:25:31,090 --> 00:25:29,600
case so do you know so we have methods

594
00:25:32,890 --> 00:25:31,100
we have we have a computational mother

595
00:25:34,930 --> 00:25:32,900
that will take just a sequence of the

596
00:25:36,910 --> 00:25:34,940
protein or take the sequence of gene

597
00:25:39,250 --> 00:25:36,920
which which can be made the genetic code

598
00:25:42,010 --> 00:25:39,260
is being very well resolved so once you

599
00:25:43,300 --> 00:25:42,020
have the journey begin you can fall you

600
00:25:45,370 --> 00:25:43,310
can figure out what the protein looks

601
00:25:47,470 --> 00:25:45,380
like and then from the protein you can

602
00:25:50,200 --> 00:25:47,480
figure out what how it will fall like

603
00:25:52,510 --> 00:25:50,210
using our methods including David bakers

604
00:25:55,900 --> 00:25:52,520
right now we can pretty structures on

605
00:25:57,460 --> 00:25:55,910
average to about 36 sanctions I would

606
00:25:59,610 --> 00:25:57,470
actually say we can do this for seventy

607
00:26:02,770 --> 00:25:59,620
percent of the proteins that are

608
00:26:04,900 --> 00:26:02,780
amenable to experiment by x-ray

609
00:26:07,900 --> 00:26:04,910
diffraction and a mass spectroscopy and

610
00:26:09,730 --> 00:26:07,910
that's an important point that that that

611
00:26:11,800 --> 00:26:09,740
there are a lot of proteins out in the

612
00:26:14,170 --> 00:26:11,810
universe that are not available to these

613
00:26:17,890 --> 00:26:14,180

two extra techniques and I'm not a woman

614
00:26:19,840 --> 00:26:17,900
going to delve into that subject but but

615
00:26:22,450 --> 00:26:19,850
anything we have these competitions

616
00:26:23,330 --> 00:26:22,460
every two years that measure how well we

617
00:26:25,580 --> 00:26:23,340
do in

618
00:26:27,230 --> 00:26:25,590
these individuals so we give and I'll

619
00:26:31,670 --> 00:26:27,240
tell you a little bit about that in a

620
00:26:35,060 --> 00:26:31,680
minute but the exposed techniques

621
00:26:36,650 --> 00:26:35,070
actually do not cover all proteins they

622
00:26:39,650 --> 00:26:36,660
do not cover membrane proteins we are

623
00:26:41,540 --> 00:26:39,660
talking about very simple in a sense a

624
00:26:43,460 --> 00:26:41,550
simple simple protein that soluble

625
00:26:45,530 --> 00:26:43,470
globular it falls nicely and

626
00:26:48,200 --> 00:26:45,540
well-behaved is what i call it like to

627
00:26:51,580 --> 00:26:48,210
like to call it and then there's

628
00:26:53,930 --> 00:26:51,590
homology so like I said I've

629
00:26:56,150 --> 00:26:53,940
evolutionary uses things again and again

630
00:26:58,760 --> 00:26:56,160
so if you take advantage of that fact

631
00:27:01,190 --> 00:26:58,770
that evolutions use structure again and

632
00:27:04,190 --> 00:27:01,200
again you can do something called

633
00:27:06,500 --> 00:27:04,200
homogeneous or company I mean

634
00:27:08,390 --> 00:27:06,510
comparative modeling and it's simply

635
00:27:09,860 --> 00:27:08,400
based so you have the structure that has

636
00:27:12,920 --> 00:27:09,870
been solved experimentally and use that

637
00:27:14,930 --> 00:27:12,930
as your basis for modeling the protein

638
00:27:18,920 --> 00:27:14,940

of something that you don't know the

639
00:27:21,740 --> 00:27:18,930
answer and those can tell those those

640
00:27:27,350 --> 00:27:21,750
cans Rygel exponential accuracy in some

641
00:27:30,890 --> 00:27:27,360
cases so very small fraction of proteins

642
00:27:34,280 --> 00:27:30,900
are small as 100 the average sized

643
00:27:36,200 --> 00:27:34,290
average domain size so let's say the

644
00:27:42,170 --> 00:27:36,210
domain is a functional unit of a protein

645
00:27:44,900 --> 00:27:42,180
there is domain sizes about 200 so when

646
00:27:47,060 --> 00:27:44,910
I say is there are 600 unique folds

647
00:27:50,750 --> 00:27:47,070
unique shapes and talk longer domains

648
00:27:52,610 --> 00:27:50,760
and I'm talking about unique chains that

649
00:27:54,020 --> 00:27:52,620
that you can splice the rest of the

650
00:27:57,410 --> 00:27:54,030
Prairie enough and they will still fold

651
00:27:59,930 --> 00:27:57,420
up into a shape and there's about six

652
00:28:03,020 --> 00:27:59,940
hundred of them versus an initiate now

653
00:28:04,670 --> 00:28:03,030
now protein can be composed of it can be

654
00:28:08,290 --> 00:28:04,680
a long chain that's composed of like two

655
00:28:11,090 --> 00:28:08,300
three domains or two three shapes but

656
00:28:14,210 --> 00:28:11,100
yeah it is there is about six or of them

657
00:28:16,790 --> 00:28:14,220
and we can get up to 150 right now we

658
00:28:18,170 --> 00:28:16,800
could probably even do 200 if we push to

659
00:28:20,470 --> 00:28:18,180
do it we it's just a matter of

660
00:28:25,970 --> 00:28:20,480
computational Thornton Universal dessert

661
00:28:28,060 --> 00:28:25,980
no 200 immense average yeah the average

662
00:28:33,400 --> 00:28:28,070
domain size external domain essence of

663
00:28:35,480 --> 00:28:33,410

real proteins in nature and 200 times

664
00:28:37,960 --> 00:28:35,490
three or four hundred

665
00:28:42,710 --> 00:28:37,970
is the number mineral acids that your

666
00:28:44,650 --> 00:28:42,720
simulated 200 times I'm sorry two of the

667
00:28:48,110 --> 00:28:44,660
200-meter acids in the domain a monster

668
00:28:52,340 --> 00:28:48,120
yeah and you said 100 domains you're

669
00:28:55,030 --> 00:28:52,350
doing no no i adn't a number 100 i I

670
00:28:59,090 --> 00:28:55,040
mean I said 200 or minor assets and then

671
00:29:02,210 --> 00:28:59,100
we can we can we can there are the 600

672
00:29:04,549 --> 00:29:02,220
unique domains and for a given protein

673
00:29:07,520 --> 00:29:04,559
we don't know the answer to that it can

674
00:29:09,080 --> 00:29:07,530
be with the G go on with it the method

675
00:29:11,930 --> 00:29:09,090
is called you know for that reason we

676
00:29:14,000 --> 00:29:11,940
just give me the sequence and we can go

677
00:29:16,730 --> 00:29:14,010
up to 150 that's pretty much it right

678
00:29:19,880 --> 00:29:16,740
now we can't do 200 so we can't even hit

679
00:29:21,799 --> 00:29:19,890
the average yet but but we've pushed it

680
00:29:28,160 --> 00:29:21,809
it's enough of computational part it's

681
00:29:31,549 --> 00:29:28,170
not a matter of really the physics but

682
00:29:34,850 --> 00:29:31,559
we did with the comparative method we

683
00:29:36,710 --> 00:29:34,860
can go any any event we can go thousands

684
00:29:39,730 --> 00:29:36,720
of miles in fact our mortality thousand

685
00:29:43,700 --> 00:29:39,740
oh manasa protein so that's you know

686
00:29:46,940 --> 00:29:43,710
3,000 times 10 which is yeah 30,000

687
00:29:48,680 --> 00:29:46,950
atoms and then model that to like two

688
00:29:51,740 --> 00:29:48,690

angstroms or three actions are nasty

689
00:29:54,430 --> 00:29:51,750
from the right real answer so with the

690
00:29:57,020 --> 00:29:54,440
reusing homology or evolution as a guide

691
00:30:00,290 --> 00:29:57,030
we can we can really go very far with a

692
00:30:04,490 --> 00:30:00,300
modeling process but in a sense we are

693
00:30:05,900 --> 00:30:04,500
starting from some something and we're

694
00:30:08,570 --> 00:30:05,910
trying to make it move towards the right

695
00:30:10,400 --> 00:30:08,580
answer and that bad refinement problem

696
00:30:12,440 --> 00:30:10,410
for the first time has been addressed

697
00:30:15,169 --> 00:30:12,450
two years ago it's a still unsolved

698
00:30:19,970 --> 00:30:15,179
problem and then there are hybrid method

699
00:30:23,690 --> 00:30:19,980
so most proteins i would say i would say

700
00:30:25,610 --> 00:30:23,700
right now I mean this is again where you

701
00:30:28,940 --> 00:30:25,620
probably buy a new news to most people

702
00:30:32,450 --> 00:30:28,950
as well but i would say that seventy

703
00:30:36,470 --> 00:30:32,460
percent of proteins in the universe or

704
00:30:40,070 --> 00:30:36,480
in the human orsa or in the human or in

705
00:30:42,710 --> 00:30:40,080
any organism are not amenable to these

706
00:30:45,140 --> 00:30:42,720
experience techniques a lot of the

707
00:30:47,560 --> 00:30:45,150
membrane-bound so exchange a mark on

708
00:30:49,300 --> 00:30:47,570
even get to that so to do exhale

709
00:30:51,310 --> 00:30:49,310
experiment in death you know nice

710
00:30:54,850 --> 00:30:51,320
crystallization India very well-behaved

711
00:30:56,590 --> 00:30:54,860
proteins so I I right now estimated you

712
00:30:58,750 --> 00:30:56,600
know when I started doing this work I

713
00:31:00,790 --> 00:30:58,760

thought it was ninety percent that that

714
00:31:02,350 --> 00:31:00,800
most proteins well behaved and now I

715
00:31:03,940 --> 00:31:02,360
think only thirty percent of proteins

716
00:31:06,070 --> 00:31:03,950
are well be here they need the

717
00:31:10,720 --> 00:31:06,080
environment to fold it they need the

718
00:31:14,200 --> 00:31:10,730
rest of the context to do to form the

719
00:31:16,720 --> 00:31:14,210
structure and so we use hybrid technics

720
00:31:18,580 --> 00:31:16,730
so we take some data for an excellent we

721
00:31:22,510 --> 00:31:18,590
combine into the computational methods

722
00:31:24,490 --> 00:31:22,520
and then and then we do and I mean again

723
00:31:27,520 --> 00:31:24,500
I can go to very much deal to do this

724
00:31:31,150 --> 00:31:27,530
but we do a hybrid simulation and that

725
00:31:33,340 --> 00:31:31,160
turns out again produce results are as

726
00:31:38,380 --> 00:31:33,350
accurate all muscles accurate as

727
00:31:44,860 --> 00:31:38,390
experiment and when I say those things

728
00:31:46,240 --> 00:31:44,870
the mean number okay actually yeah one

729
00:31:50,020 --> 00:31:46,250
of the points I want to make right there

730
00:31:51,340 --> 00:31:50,030
is is that because this is a basis of

731
00:31:54,160 --> 00:31:51,350
all our methods that we're developing

732
00:31:56,350 --> 00:31:54,170
and it's in it's an algorithmic issue

733
00:31:58,510 --> 00:31:56,360
but one of the things that the more

734
00:32:00,040 --> 00:31:58,520
distance constraints that we have or the

735
00:32:03,220 --> 00:32:00,050
more we know about distances between

736
00:32:06,640 --> 00:32:03,230
atoms the more we can specify what the

737
00:32:10,030 --> 00:32:06,650
structure looks like and as you add more

738
00:32:11,740 --> 00:32:10,040

and more distance constraints you can

739
00:32:16,360 --> 00:32:11,750
get to the structure and this is being

740
00:32:18,100 --> 00:32:16,370
published in many places but but we've

741
00:32:19,960 --> 00:32:18,110
shown that you know one business concern

742
00:32:21,700 --> 00:32:19,970
for every ten of mine assets that is to

743
00:32:24,760 --> 00:32:21,710
domain assets pretty very far away from

744
00:32:26,920 --> 00:32:24,770
each other can get your low resolution

745
00:32:28,750 --> 00:32:26,930
we call this a low resolution for and

746
00:32:31,930 --> 00:32:28,760
when distance consent for every six

747
00:32:36,220 --> 00:32:31,940
amino acids six residues because rescue

748
00:32:37,870 --> 00:32:36,230
is part of a protein then we we can get

749
00:32:40,390 --> 00:32:37,880
something that matches experimental

750
00:32:45,880 --> 00:32:40,400
accuracy so that's the point that that

751
00:32:56,390 --> 00:32:52,400
and okay so why is this such a hard

752
00:32:59,570 --> 00:32:56,400
cutter why is this so hard okay so let's

753
00:33:02,510 --> 00:32:59,580
let's let's go back to that and so did

754
00:33:04,850 --> 00:33:02,520
normal prediction so the idea is just we

755
00:33:06,080 --> 00:33:04,860
are approaches supple conformational

756
00:33:08,330 --> 00:33:06,090
space such that needle-like

757
00:33:10,910 --> 00:33:08,340
confirmations are found so if I go back

758
00:33:14,180 --> 00:33:10,920
right here to that slide actually it's

759
00:33:16,550 --> 00:33:14,190
already there but if I go back here

760
00:33:18,830 --> 00:33:16,560
right here I show you this unfolded

761
00:33:20,060 --> 00:33:18,840
protein it has degrees of freedom and

762
00:33:22,580 --> 00:33:20,070
there's actually two degrees of freedom

763
00:33:25,820 --> 00:33:22,590

we call the fire and Phi angle they can

764
00:33:27,770 --> 00:33:25,830
twist around 360 degrees and they can't

765
00:33:30,020 --> 00:33:27,780
really to some turn 60 degrees because

766
00:33:32,060 --> 00:33:30,030
their other atoms interfering with it so

767
00:33:34,550 --> 00:33:32,070
we know what the distributions that they

768
00:33:44,029 --> 00:33:34,560
can go around and we use that as part of

769
00:33:50,599 --> 00:33:48,979
yeah so the our approach is to I'll

770
00:34:01,799 --> 00:33:50,609
actually he'll actually skip a lot of

771
00:34:07,960 --> 00:34:05,799
yeah our approaches to basically

772
00:34:09,309 --> 00:34:07,970
examples this is a huge confirmation

773
00:34:11,440 --> 00:34:09,319
spend anything but it's an actually

774
00:34:14,440 --> 00:34:11,450
infinite conformational space although

775
00:34:16,329 --> 00:34:14,450
all 360 angles and all real numbers but

776
00:34:19,629 --> 00:34:16,339
even if you say that there are five

777
00:34:22,539 --> 00:34:19,639
states / protein and sorry for a man

778
00:34:24,549 --> 00:34:22,549
acid and you although you have a small

779
00:34:27,069 --> 00:34:24,559
protein 100 and minor assets which is

780
00:34:31,089 --> 00:34:27,079
small domain as they just said that's

781
00:34:34,859 --> 00:34:31,099
five to one hundred poverty's which is

782
00:34:38,049 --> 00:34:34,869
more than what we can look at and and

783
00:34:40,599 --> 00:34:38,059
that's equal to 10 to 70 and you know

784
00:34:42,549 --> 00:34:40,609
you're you guys are astrobiologist and

785
00:34:44,049 --> 00:34:42,559
you correct me if I'm wrong and this but

786
00:34:45,849 --> 00:34:44,059
the last estimate I heard about the

787
00:34:50,049 --> 00:34:45,859
number of atoms in universe was about 10

788
00:34:52,629 --> 00:34:50,059

to the 69 so so this is larger than

789
00:34:55,569 --> 00:34:52,639
number of atoms in the English so there

790
00:34:57,520 --> 00:34:55,579
is no way that the protein has can can

791
00:35:00,460 --> 00:34:57,530
sample all of this thing what has

792
00:35:02,650 --> 00:35:00,470
happened again is evolution right over

793
00:35:06,599 --> 00:35:02,660
time billions of years this this thing

794
00:35:10,120 --> 00:35:06,609
is evolved to to get to its right shape

795
00:35:11,770 --> 00:35:10,130
to perform at right function and the

796
00:35:13,599 --> 00:35:11,780
things the organisms that don't perform

797
00:35:15,250 --> 00:35:13,609
the rite functions die out and the

798
00:35:18,220 --> 00:35:15,260
organism that performs that I functions

799
00:35:20,140 --> 00:35:18,230
arrive and so on so evolution has helped

800
00:35:22,660 --> 00:35:20,150
guide this process and we are trying to

801
00:35:24,579 --> 00:35:22,670
replicate that in some ways so you have

802
00:35:27,160 --> 00:35:24,589
a huge confirmation space and be cutting

803
00:35:30,160 --> 00:35:27,170
a look at it because so large so we

804
00:35:32,559 --> 00:35:30,170
sample it wissam played using variety of

805
00:35:34,210 --> 00:35:32,569
energy functions and we hope that within

806
00:35:37,329 --> 00:35:34,220
our sample there's something that looks

807
00:35:39,789 --> 00:35:37,339
like the real answer and then the second

808
00:35:43,089 --> 00:35:39,799
hard problem is to figure out which one

809
00:35:44,530 --> 00:35:43,099
it is that's the search and it's very

810
00:35:46,839 --> 00:35:44,540
very very hard when you have such a

811
00:35:49,720 --> 00:35:46,849
large sample size we right now look at

812
00:35:53,200 --> 00:35:49,730
10 to the 11 10 to the 12 confirmations

813
00:35:55,690 --> 00:35:53,210

so more than a billion easy all the Onyx

814
00:35:58,089 --> 00:35:55,700
it's very very hard to design it

815
00:36:00,039 --> 00:35:58,099
something a function or someone will say

816
00:36:03,309 --> 00:36:00,049
hey this is the right answer what's all

817
00:36:04,839 --> 00:36:03,319
the others and so these are this is the

818
00:36:06,520 --> 00:36:04,849
reason why the structure prediction

819
00:36:11,109 --> 00:36:06,530
problem the protein folding problem is

820
00:36:13,390 --> 00:36:11,119
so hard and the technique for doing it I

821
00:36:14,089 --> 00:36:13,400
will not get any calendar template must

822
00:36:16,009 --> 00:36:14,099
be smart

823
00:36:18,979 --> 00:36:16,019
want to see a little bit about it in the

824
00:36:21,819 --> 00:36:18,989
sense that what we do is that that's the

825
00:36:24,349 --> 00:36:21,829
method some people in the method because

826
00:36:27,710 --> 00:36:24,359
because it requires all knowledge about

827
00:36:29,630 --> 00:36:27,720
how we do the simulations skin elected

828
00:36:30,950 --> 00:36:29,640
the duplex method based method is taking

829
00:36:33,559 --> 00:36:30,960
a protein in comparing it to the

830
00:36:36,160 --> 00:36:33,569
database of known structure 45,000 about

831
00:36:39,249 --> 00:36:36,170
45,000 known structures that have been

832
00:36:43,009 --> 00:36:39,259
solved bikes a diffraction or by

833
00:36:45,950 --> 00:36:43,019
crystallography and we want to we want

834
00:36:48,289 --> 00:36:45,960
to use that knowledge so when we detect

835
00:36:50,599 --> 00:36:48,299
a homology or similarity then we can use

836
00:36:52,999 --> 00:36:50,609
the alignment between them and we can

837
00:36:55,219 --> 00:36:53,009
come up with an initial model and then

838
00:36:57,739 --> 00:36:55,229

we can we can we can actually use that

839
00:37:00,769 --> 00:36:57,749
initial model again going very fast on

840
00:37:03,440 --> 00:37:00,779
this this part of it they can use that

841
00:37:08,120 --> 00:37:03,450
as a template for guiding what our

842
00:37:09,710 --> 00:37:08,130
structure will finally look like and we

843
00:37:11,539 --> 00:37:09,720
have methods to do that you have a lot

844
00:37:16,479 --> 00:37:11,549
of techniques as spent 14 years like a

845
00:37:18,799 --> 00:37:16,489
set of my life doing this myself and and

846
00:37:19,969 --> 00:37:18,809
the main thing is I have any concept of

847
00:37:22,910 --> 00:37:19,979
someone's one of this thing actually

848
00:37:24,650 --> 00:37:22,920
looks like that and you want to do it to

849
00:37:26,749 --> 00:37:24,660
what the serial answer looks like and

850
00:37:29,329 --> 00:37:26,759
it's very hard to that right downtown

851
00:37:31,130 --> 00:37:29,339
saw problem until last year where we

852
00:37:33,499 --> 00:37:31,140
were when the first people to do that

853
00:37:35,150 --> 00:37:33,509
where we could actually move something

854
00:37:36,920 --> 00:37:35,160
that was like say three angstroms away

855
00:37:39,049 --> 00:37:36,930
from the car translated to angle the way

856
00:37:41,599 --> 00:37:39,059
to the real answer so I want to give you

857
00:37:45,109 --> 00:37:41,609
some results now actually I always keep

858
00:37:47,180 --> 00:37:45,119
all method slides but um so we get

859
00:37:49,219 --> 00:37:47,190
assessed every two years where this

860
00:37:53,559 --> 00:37:49,229
competition has become a competition is

861
00:37:56,359 --> 00:37:53,569
called cash and what happens is that the

862
00:38:00,109 --> 00:37:56,369
modelers get sequences of proteins that

863
00:38:02,450 --> 00:38:00,119

are not being published or hair in the

864
00:38:04,609 --> 00:38:02,460
process of being solved so so they're

865
00:38:06,739 --> 00:38:04,619
about to be solved so we get all the

866
00:38:09,620 --> 00:38:06,749
sequences in saying may April or May

867
00:38:12,950 --> 00:38:09,630
during we called the casp season and

868
00:38:15,410 --> 00:38:12,960
then and then and then the

869
00:38:16,940 --> 00:38:15,420
crystallographers NMR spectra is working

870
00:38:19,749 --> 00:38:16,950
very hard and getting the right answer

871
00:38:22,910 --> 00:38:19,759
and we try to predict the structure and

872
00:38:24,589 --> 00:38:22,920
then this is buying prediction what has

873
00:38:25,910 --> 00:38:24,599
happened in the past is that people have

874
00:38:27,030 --> 00:38:25,920
claimed to have solve this problem many

875
00:38:30,420 --> 00:38:27,040
many many many

876
00:38:33,030 --> 00:38:30,430
so the literature is filled with my

877
00:38:36,380 --> 00:38:33,040
mentor is actually you know it's a big

878
00:38:38,910 --> 00:38:36,390
job but but the roof is filled with

879
00:38:40,590 --> 00:38:38,920
people who have said that you know they

880
00:38:42,420 --> 00:38:40,600
take a stair set of ten proteins and

881
00:38:45,150 --> 00:38:42,430
they write a program to work on it and

882
00:38:46,740 --> 00:38:45,160
it usually fails to try to fold the

883
00:38:48,300 --> 00:38:46,750
structure and then they treat their

884
00:38:50,550 --> 00:38:48,310
program to work make it work better and

885
00:38:52,470 --> 00:38:50,560
by doing so they introduce knowledge

886
00:38:54,810 --> 00:38:52,480
about the test set into their algorithm

887
00:38:57,210 --> 00:38:54,820
and they keep doing that and finally it

888
00:38:59,040 --> 00:38:57,220

really works very well on their ten test

889
00:39:00,120 --> 00:38:59,050
proteins that they're looking at but

890
00:39:02,790 --> 00:39:00,130
then when they're given an unknown

891
00:39:05,550 --> 00:39:02,800
protein there they are not they're not

892
00:39:07,200 --> 00:39:05,560
able to do so well so John well my first

893
00:39:08,790 --> 00:39:07,210
mentor came up with the idea of doing

894
00:39:11,250 --> 00:39:08,800
this structure prediction of blind way

895
00:39:13,980 --> 00:39:11,260
so this is minus cos gasp and this is a

896
00:39:15,660 --> 00:39:13,990
competition and I'm talking about so we

897
00:39:17,220 --> 00:39:15,670
do very well length in these rankings

898
00:39:19,020 --> 00:39:17,230
and so on so what I'm showing you on the

899
00:39:21,390 --> 00:39:19,030
left is a real answer and what I'm

900
00:39:24,030 --> 00:39:21,400
showing on the right is the model that

901
00:39:26,130 --> 00:39:24,040
we produced so here when you have

902
00:39:28,260 --> 00:39:26,140
something at a 60-person similar that is

903
00:39:30,060 --> 00:39:28,270
not not sixty percent similar sixty

904
00:39:32,400 --> 00:39:30,070
percent identical it does not matter

905
00:39:33,960 --> 00:39:32,410
acids you can produce something that's

906
00:39:35,940 --> 00:39:33,970
almost as good as experimental

907
00:39:37,260 --> 00:39:35,950
resolution that means that if you took

908
00:39:39,300 --> 00:39:37,270
the structure and saw it in two

909
00:39:41,480 --> 00:39:39,310
different labs you will probably get

910
00:39:45,660 --> 00:39:41,490
something that that is about this much

911
00:39:47,790 --> 00:39:45,670
CFR masti but we are not interested in

912
00:39:49,680 --> 00:39:47,800
the sixty percent well we are but but

913
00:39:52,140 --> 00:39:49,690

those are the easy ones let's go into

914
00:39:54,240 --> 00:39:52,150
twenty five percent then there's a lot

915
00:39:55,710 --> 00:39:54,250
of divergence in sequence but the

916
00:39:57,420 --> 00:39:55,720
structure is like I said I'm more

917
00:39:59,850 --> 00:39:57,430
concerned in sequence here's another

918
00:40:02,190 --> 00:39:59,860
prediction where we detect that and we

919
00:40:05,210 --> 00:40:02,200
say hey look this this actually is this

920
00:40:08,160 --> 00:40:05,220
thing and and here we get 2.2 anxious

921
00:40:10,940 --> 00:40:08,170
prediction for something that's a

922
00:40:15,110 --> 00:40:10,950
five-person similar here's another one

923
00:40:17,970 --> 00:40:15,120
2.0 action for something that's

924
00:40:21,030 --> 00:40:17,980
twenty-three percent similar and here's

925
00:40:22,560 --> 00:40:21,040
another result that it's actually 11

926
00:40:26,340 --> 00:40:22,570
person that's all of these by the way

927
00:40:29,600 --> 00:40:26,350
anything about about 22 and below are

928
00:40:32,040 --> 00:40:29,610
the random range so you could get these

929
00:40:35,720 --> 00:40:32,050
hits by chance when you do a side

930
00:40:38,250 --> 00:40:35,730
blasters that's what most people do and

931
00:40:39,290 --> 00:40:38,260
so a loving person is definitely a

932
00:40:40,880 --> 00:40:39,300
random

933
00:40:42,710 --> 00:40:40,890
headbutt you and then these two proteins

934
00:40:44,510 --> 00:40:42,720
are structurally related if you actually

935
00:40:47,660 --> 00:40:44,520
follow c alpha traces you can see that

936
00:40:49,370 --> 00:40:47,670
their ships are kind of similar and the

937
00:40:51,470 --> 00:40:49,380
question is can you move this back to

938
00:40:53,960 --> 00:40:51,480

the real answer and we're working on

939
00:40:55,960 --> 00:40:53,970
that i'm showing you four examples here

940
00:40:59,240 --> 00:40:55,970
we usually model about 100 proteins

941
00:41:01,280 --> 00:40:59,250
doing the casp season and we get judged

942
00:41:05,270 --> 00:41:01,290
on that and we get ranked on that and

943
00:41:06,680 --> 00:41:05,280
and well I won't tell you what my

944
00:41:09,770 --> 00:41:06,690
rankings would be but but it's in the

945
00:41:25,890 --> 00:41:09,780
top five let's put it that way it

946
00:41:25,900 --> 00:41:29,260
you

947
00:41:38,050 --> 00:41:34,690
sorry eyeing some problems this I'm not

948
00:41:40,060 --> 00:41:38,060
used to make okay so given a structure

949
00:41:41,380 --> 00:41:40,070
we want to break the function and this

950
00:41:44,410 --> 00:41:41,390
is where some of the astrological

951
00:41:48,220 --> 00:41:44,420
aspects come in so what we've developed

952
00:41:50,500 --> 00:41:48,230
is is a function so I'm in the

953
00:41:52,780 --> 00:41:50,510
Department of Microbiology and they had

954
00:41:54,730 --> 00:41:52,790
me I mean I'm glad they had me because I

955
00:41:56,650 --> 00:41:54,740
personally there were a lot of

956
00:41:58,930 --> 00:41:56,660
biochemistry and biophysics departments

957
00:42:02,110 --> 00:41:58,940
that wanted to hire me all over the

958
00:42:04,030 --> 00:42:02,120
country and I chose microbiology at the

959
00:42:06,070 --> 00:42:04,040
other because I thought they would

960
00:42:07,540 --> 00:42:06,080
challenge me and they did they say okay

961
00:42:10,990 --> 00:42:07,550
so you give me the structure so what

962
00:42:12,460 --> 00:42:11,000
what can I do with it and so we give you

963
00:42:14,530 --> 00:42:12,470

a winter structure we're going to the

964
00:42:16,210 --> 00:42:14,540
function because we believe that the

965
00:42:17,980 --> 00:42:16,220
structure determines function I mean

966
00:42:20,110 --> 00:42:17,990
that's a fundamental rule so we want to

967
00:42:22,360 --> 00:42:20,120
get to get our function so now we

968
00:42:24,490 --> 00:42:22,370
started developing new scoring functions

969
00:42:27,400 --> 00:42:24,500
that are in this party working on in

970
00:42:30,420 --> 00:42:27,410
fact to try to predict the function of

971
00:42:34,000 --> 00:42:30,430
what the port of given protein does and

972
00:42:35,830 --> 00:42:34,010
so we got this function this is the

973
00:42:37,180 --> 00:42:35,840
earlier version of squaring function and

974
00:42:40,240 --> 00:42:37,190
actually is a correlation coefficient of

975
00:42:45,370 --> 00:42:40,250
point 7 to explain lida Turman binding

976
00:42:48,190 --> 00:42:45,380
affinity to some metal ion 2 2 the score

977
00:42:51,120 --> 00:42:48,200
that we predict so that's pretty good

978
00:42:53,710 --> 00:42:51,130
that means that we can start breaking

979
00:42:56,530 --> 00:42:53,720
proteins of lions actually ions in

980
00:42:58,060 --> 00:42:56,540
general in crystal structures or no more

981
00:42:59,500 --> 00:42:58,070
structures don't have a resolution you

982
00:43:03,100 --> 00:42:59,510
can't resolve them what kind of ions

983
00:43:05,530 --> 00:43:03,110
they are or where they are and we can

984
00:43:08,500 --> 00:43:05,540
actually say that we can get them here

985
00:43:10,450 --> 00:43:08,510
i'm showing four examples again so here

986
00:43:13,000 --> 00:43:10,460
the yellow is actually doesn't matter

987
00:43:15,220 --> 00:43:13,010
the yellow is the correct answer and the

988
00:43:18,220 --> 00:43:15,230

blue is our prediction and they're all a

989
00:43:21,820 --> 00:43:18,230
pain so we are predicting calcium ions

990
00:43:23,890 --> 00:43:21,830
in proteins also calcium is a big

991
00:43:26,980 --> 00:43:23,900
regulator of most most biological

992
00:43:28,630 --> 00:43:26,990
functions I had epilepsy for example and

993
00:43:30,460 --> 00:43:28,640
I can tell you that that there's a

994
00:43:33,670 --> 00:43:30,470
deficient in calcium every time I have

995
00:43:38,020 --> 00:43:33,680
it I have a seizure and so so we can

996
00:43:41,470 --> 00:43:38,030
prick the the the the accuracy of

997
00:43:42,160 --> 00:43:41,480
calcium ions 2.05 our angstroms rms you

998
00:43:44,500 --> 00:43:42,170
where you

999
00:43:46,000 --> 00:43:44,510
very accurately so that mean that's what

1000
00:43:48,880 --> 00:43:46,010
i mean by that mean we're on top of

1001
00:43:51,880 --> 00:43:48,890
pretty much on top and 103 test cases

1002
00:43:52,840 --> 00:43:51,890
that we looked at and then we are

1003
00:43:55,840 --> 00:43:52,850
looking at something called meta

1004
00:43:58,000 --> 00:43:55,850
functional signature so this is more of

1005
00:44:01,750 --> 00:43:58,010
a bias physics based approach but this

1006
00:44:03,850 --> 00:44:01,760
is combining sequence so most colleges

1007
00:44:05,560 --> 00:44:03,860
will only look at this column right here

1008
00:44:08,590 --> 00:44:05,570
they look at sequence conservation and

1009
00:44:10,480 --> 00:44:08,600
their family of sequences and say hey

1010
00:44:12,070 --> 00:44:10,490
this looks this is comes out throughout

1011
00:44:16,360 --> 00:44:12,080
this whole family so this must be

1012
00:44:17,800 --> 00:44:16,370
important but but then what happens if

1013
00:44:20,020 --> 00:44:17,810

you have to evolutionary lineages that

1014
00:44:22,030 --> 00:44:20,030
I've drip diverged you have plants on

1015
00:44:23,620 --> 00:44:22,040
your animals and it's concerned or more

1016
00:44:26,380 --> 00:44:23,630
plants and it's not conserved among and

1017
00:44:28,240 --> 00:44:26,390
animals so in the plants it does

1018
00:44:29,680 --> 00:44:28,250
probably probably does do the function

1019
00:44:31,900 --> 00:44:29,690
but it animals maybe it's not so

1020
00:44:33,850 --> 00:44:31,910
important anymore so we take that into

1021
00:44:36,850 --> 00:44:33,860
account and then we take the structure

1022
00:44:39,940 --> 00:44:36,860
of the stack we take the structuring car

1023
00:44:43,720 --> 00:44:39,950
one of the things that that that from a

1024
00:44:46,570 --> 00:44:43,730
physical point of view is that anything

1025
00:44:49,210 --> 00:44:46,580
that needs to do a function is not

1026
00:44:51,340 --> 00:44:49,220
structurally stable so anything that

1027
00:44:55,750 --> 00:44:51,350
that is important for structure has to

1028
00:44:58,020 --> 00:44:55,760
be important for function so if you put

1029
00:45:00,970 --> 00:44:58,030
energy in the middle of a protein and

1030
00:45:02,890 --> 00:45:00,980
destabilizes of protein it will unfold

1031
00:45:04,540 --> 00:45:02,900
and it will cause a loss of function

1032
00:45:06,730 --> 00:45:04,550
because it completely destroys the

1033
00:45:08,920 --> 00:45:06,740
protein so anything is structurally

1034
00:45:10,570 --> 00:45:08,930
important for protein is important for

1035
00:45:12,450 --> 00:45:10,580
the function but the reverse doesn't

1036
00:45:14,710 --> 00:45:12,460
hold anything that is functionally

1037
00:45:16,840 --> 00:45:14,720
important is actually the other way

1038
00:45:18,430 --> 00:45:16,850

around it's functionally frustrated so

1039
00:45:21,070 --> 00:45:18,440
you publish a paper on P and X on this

1040
00:45:23,770 --> 00:45:21,080
where we show that when the function

1041
00:45:25,360 --> 00:45:23,780
actually occurs then it's happy but

1042
00:45:27,730 --> 00:45:25,370
until then it's actually not happy that

1043
00:45:29,170 --> 00:45:27,740
I'm an acid or the minus is involved in

1044
00:45:31,420 --> 00:45:29,180
this its weak i recalled functional

1045
00:45:33,490 --> 00:45:31,430
frustration so we use that into account

1046
00:45:36,130 --> 00:45:33,500
in our in our stability measurements and

1047
00:45:38,050 --> 00:45:36,140
the energy function that we're using are

1048
00:45:39,490 --> 00:45:38,060
based on the same principles that we

1049
00:45:41,890 --> 00:45:39,500
used to do the protein structure

1050
00:45:43,630 --> 00:45:41,900
prediction so and then we can see that

1051
00:45:49,060 --> 00:45:43,640
other thousand test cases are we looking

1052
00:45:51,859 --> 00:45:49,070
at we're getting me a hundred almost

1053
00:45:54,400 --> 00:45:51,869
hundred percent prediction on predicting

1054
00:45:58,160 --> 00:45:54,410
which we chose use a function important

1055
00:45:59,660 --> 00:45:58,170
and this is in fact i want to i want to

1056
00:46:03,109 --> 00:45:59,670
bring at this point this is what Aaron's

1057
00:46:06,799 --> 00:46:03,119
actually developing and working on and

1058
00:46:09,589 --> 00:46:06,809
improving and it has to relevance s 2 to

1059
00:46:11,569 --> 00:46:09,599
s herbology one is that you know from my

1060
00:46:13,819 --> 00:46:11,579
perspective on my land biologist or a

1061
00:46:15,529 --> 00:46:13,829
competition largest I want to know how

1062
00:46:17,900 --> 00:46:15,539
things by and I want to design new drugs

1063
00:46:20,420 --> 00:46:17,910

and things like that but we can design

1064
00:46:22,609 --> 00:46:20,430
new proteins with this we can design

1065
00:46:24,559 --> 00:46:22,619
functionally new proteins that have the

1066
00:46:27,109 --> 00:46:24,569
properties that we won based on the

1067
00:46:29,509 --> 00:46:27,119
functional signatures and if you want

1068
00:46:31,160 --> 00:46:29,519
the seed life on the planet say you look

1069
00:46:35,059 --> 00:46:31,170
at extreme environments on this on this

1070
00:46:37,309 --> 00:46:35,069
on this planet and say you know these

1071
00:46:39,440 --> 00:46:37,319
thermal files are these yeah this

1072
00:46:43,309 --> 00:46:39,450
thermophilic bacteria actually errands

1073
00:46:45,140 --> 00:46:43,319
looked at this do a particular function

1074
00:46:46,910 --> 00:46:45,150
and that they actually differ in this

1075
00:46:49,099 --> 00:46:46,920
function signature and we can replicate

1076
00:46:51,559 --> 00:46:49,109
that aspect of it might still keep it

1077
00:46:53,660 --> 00:46:51,569
structurally stable so we use that the

1078
00:46:55,099 --> 00:46:53,670
same same approach and we use the

1079
00:46:58,190 --> 00:46:55,109
structural prediction with the

1080
00:47:00,039 --> 00:46:58,200
functional collection to to to try to

1081
00:47:02,630 --> 00:47:00,049
design new proteins and that's an

1082
00:47:07,849 --> 00:47:02,640
astrological application or what Aaron

1083
00:47:10,759 --> 00:47:07,859
is doing then like I said life is

1084
00:47:12,559 --> 00:47:10,769
complex right we proteins own just work

1085
00:47:15,259 --> 00:47:12,569
by themselves to work with interactions

1086
00:47:16,759 --> 00:47:15,269
this is a case where we found Cubans and

1087
00:47:19,009 --> 00:47:16,769
bacteria this was actually done with Jim

1088
00:47:22,910 --> 00:47:19,019

Staley is a paper like Oh publish tools

1089
00:47:24,979 --> 00:47:22,920
in the service department where we

1090
00:47:29,779 --> 00:47:24,989
actually predicted in structure to 2.8

1091
00:47:32,089 --> 00:47:29,789
angstrom is sorry 2.18 angstroms for the

1092
00:47:34,099 --> 00:47:32,099
monomer to point to a tensor so diamond

1093
00:47:37,150 --> 00:47:34,109
which is really a great prediction then

1094
00:47:40,489 --> 00:47:37,160
in structure before the shuck sugar salt

1095
00:47:44,180 --> 00:47:40,499
so gingerly found tubulin sand in the

1096
00:47:45,950 --> 00:47:44,190
prosthetic of bacteria and he gave me

1097
00:47:47,239 --> 00:47:45,960
these these genes and he said what does

1098
00:47:48,729 --> 00:47:47,249
it look like do you think that they

1099
00:47:52,339 --> 00:47:48,739
actually interact and I actually said

1100
00:47:54,620 --> 00:47:52,349
destruction I didn't I said actually

1101
00:47:56,420 --> 00:47:54,630
arranged right that that is my visual

1102
00:47:59,839 --> 00:47:56,430
inspection and actually if you go back

1103
00:48:00,920 --> 00:47:59,849
to the meta functional signature here it

1104
00:48:04,250 --> 00:48:00,930
actually predicts that they're really

1105
00:48:06,410 --> 00:48:04,260
interact so what was I missing

1106
00:48:08,210 --> 00:48:06,420
I was missing the ablution image of what

1107
00:48:10,730 --> 00:48:08,220
was happening so the things that I

1108
00:48:12,620 --> 00:48:10,740
thought were important changes they are

1109
00:48:14,210 --> 00:48:12,630
not important at all because you see

1110
00:48:17,630 --> 00:48:14,220
them happening all the time in other

1111
00:48:19,550 --> 00:48:17,640
eukaryotes and so on so I was so so the

1112
00:48:20,930 --> 00:48:19,560
the album the computer album was

1113
00:48:23,300 --> 00:48:20,940

actually more accurate predicting

1114
00:48:25,070 --> 00:48:23,310
whether these damn rises and the Final

1115
00:48:28,970 --> 00:48:25,080
Four microfilaments and so on and Jim is

1116
00:48:30,800 --> 00:48:28,980
pursuing this and and this was a

1117
00:48:32,360 --> 00:48:30,810
thirty-five percent identity so we

1118
00:48:34,280 --> 00:48:32,370
actually got the model right because

1119
00:48:37,010 --> 00:48:34,290
structural not right but we couldn't get

1120
00:48:38,900 --> 00:48:37,020
the function of function right but now

1121
00:48:42,680 --> 00:48:38,910
we can with the new methods that we have

1122
00:48:45,410 --> 00:48:42,690
and we can also do the same thing with

1123
00:48:46,940 --> 00:48:45,420
protein-dna interactions so here you

1124
00:48:49,130 --> 00:48:46,950
have a transcription factor boundary

1125
00:48:51,350 --> 00:48:49,140
gear for whatever we can now be

1126
00:48:55,340 --> 00:48:51,360
completely model the lack of prime I

1127
00:48:57,860 --> 00:48:55,350
hope well if you're not well I just but

1128
00:48:59,360 --> 00:48:57,870
so I'll skip that part of it but there's

1129
00:49:01,340 --> 00:48:59,370
there's a if you know about it there's

1130
00:49:04,340 --> 00:49:01,350
something called the lacquer like opera

1131
00:49:05,990 --> 00:49:04,350
opera which is used as a proto system

1132
00:49:08,060 --> 00:49:06,000
and many many organisms we can

1133
00:49:10,400 --> 00:49:08,070
completely model that at our Thomas

1134
00:49:13,580 --> 00:49:10,410
atomistic level of detail and get

1135
00:49:16,790 --> 00:49:13,590
everything right and one of the things

1136
00:49:18,440 --> 00:49:16,800
that we found is that one of the things

1137
00:49:21,140 --> 00:49:18,450
that we are that's special about what we

1138
00:49:24,080 --> 00:49:21,150

are doing is that we took dynamics into

1139
00:49:26,270 --> 00:49:24,090
our process so we let the proteins and

1140
00:49:27,710 --> 00:49:26,280
the let's say two proteins are the

1141
00:49:30,620 --> 00:49:27,720
protein in the DNA of the protein and

1142
00:49:33,160 --> 00:49:30,630
the substrate buying with and move each

1143
00:49:35,900 --> 00:49:33,170
other and then measure their scores and

1144
00:49:37,460 --> 00:49:35,910
this is a correlation with the

1145
00:49:39,080 --> 00:49:37,470
exponential binding energy and this is

1146
00:49:40,670 --> 00:49:39,090
the docking energy of the calculated

1147
00:49:44,380 --> 00:49:40,680
energy and you can see the correlation

1148
00:49:48,590 --> 00:49:44,390
coefficient is point almost pine and

1149
00:49:50,780 --> 00:49:48,600
that turns out to be if you didn't do it

1150
00:49:55,940 --> 00:49:50,790
without the dynamics it would be 0 point

1151
00:49:58,400 --> 00:49:55,950
35 that's random so proteins substrates

1152
00:50:00,680 --> 00:49:58,410
DNA all these things are constant

1153
00:50:02,540 --> 00:50:00,690
emotion and you take dynamics into

1154
00:50:05,680 --> 00:50:02,550
account when you do this modeling so

1155
00:50:09,200 --> 00:50:05,690
that's something that we're big on doing

1156
00:50:11,330 --> 00:50:09,210
so putting it all together but put all

1157
00:50:13,910 --> 00:50:11,340
the structure functions and interactions

1158
00:50:15,130 --> 00:50:13,920
ticket so now until then until now I've

1159
00:50:17,660 --> 00:50:15,140
been talking with you

1160
00:50:19,730 --> 00:50:17,670
here's a network so you start getting

1161
00:50:22,600 --> 00:50:19,740
networks you see where the interactions

1162
00:50:27,050 --> 00:50:22,610
are playing this is an example

1163
00:50:30,980 --> 00:50:27,060

interaction network tuberculosis and we

1164
00:50:33,230 --> 00:50:30,990
are looking 107 proteins with two unique

1165
00:50:34,970 --> 00:50:33,240
interactions and you can actually look

1166
00:50:36,920 --> 00:50:34,980
at it they form what we call these hubs

1167
00:50:39,250 --> 00:50:36,930
and eight notes that's that's language

1168
00:50:42,320 --> 00:50:39,260
that's being used so these hubs are

1169
00:50:44,720 --> 00:50:42,330
crucial drug targets and again if you

1170
00:50:46,580 --> 00:50:44,730
want to design a new organism that would

1171
00:50:48,110 --> 00:50:46,590
survive in the extreme environment you

1172
00:50:52,850 --> 00:50:48,120
would need something like this hot out

1173
00:50:54,680 --> 00:50:52,860
there and then the date nodes we also

1174
00:50:56,570 --> 00:50:54,690
find these articulation points to be

1175
00:50:58,970 --> 00:50:56,580
important for the survival of the

1176
00:51:02,560 --> 00:50:58,980
organism and that was just in one

1177
00:51:05,570 --> 00:51:02,570
example but formatting problem here but

1178
00:51:08,630 --> 00:51:05,580
we're looking at 26,000 protein in human

1179
00:51:10,670 --> 00:51:08,640
and we're looking at four seventeen

1180
00:51:14,110 --> 00:51:10,680
thousand of them we can predict 828

1181
00:51:16,760 --> 00:51:14,120
thousand interactions and 1 million

1182
00:51:19,100 --> 00:51:16,770
transcription regular to interaction

1183
00:51:21,590 --> 00:51:19,110
that is protein-dna interactions so and

1184
00:51:23,780 --> 00:51:21,600
like I said rice is one of my major

1185
00:51:25,610 --> 00:51:23,790
funding sources so we look at 65 streams

1186
00:51:28,760 --> 00:51:25,620
and we are trying to actually engineer

1187
00:51:32,060 --> 00:51:28,770
rice to have a whole wide of

1188
00:51:35,570 --> 00:51:32,070

bioavailable nutrients this is part of

1189
00:51:38,240 --> 00:51:35,580
the Gates Foundation I fired so that the

1190
00:51:39,650 --> 00:51:38,250
idea is that people who eat rice in most

1191
00:51:41,480 --> 00:51:39,660
Asian countries don't get all the

1192
00:51:43,430 --> 00:51:41,490
nutrients they need so can we actually

1193
00:51:45,740 --> 00:51:43,440
engineer all of this and we don't want

1194
00:51:47,810 --> 00:51:45,750
to do this through genetic modifications

1195
00:51:49,250 --> 00:51:47,820
like the Golden Rice thing because that

1196
00:51:51,260 --> 00:51:49,260
is not socially acceptable or

1197
00:51:52,730 --> 00:51:51,270
politically acceptable so you want to do

1198
00:51:53,570 --> 00:51:52,740
it through markers of screening that

1199
00:51:55,490 --> 00:51:53,580
means that you need to know the

1200
00:51:58,160 --> 00:51:55,500
functions all these networks and how

1201
00:51:59,750 --> 00:51:58,170
these networks mix together so in sum we

1202
00:52:02,180 --> 00:51:59,760
can predict function from one fifty

1203
00:52:03,230 --> 00:52:02,190
percent of proteome approximately ten

1204
00:52:05,950 --> 00:52:03,240
more day and protein-protein

1205
00:52:08,600 --> 00:52:05,960
interactions and putting linux

1206
00:52:11,990 --> 00:52:08,610
interactions that when i give you these

1207
00:52:14,420 --> 00:52:12,000
numbers we can benchmark the accuracy

1208
00:52:17,740 --> 00:52:14,430
and we've done it maybe I've got send it

1209
00:52:21,050 --> 00:52:17,750
test set and we can benchmark that and

1210
00:52:23,720 --> 00:52:21,060
you know the more accurate you get the

1211
00:52:25,160 --> 00:52:23,730
less the coverage but if you want fifty

1212
00:52:27,080 --> 00:52:25,170
percent accuracy which is what you

1213
00:52:27,870 --> 00:52:27,090

actually expect a high throughput is to

1214
00:52:32,759 --> 00:52:27,880
hybrid experiment

1215
00:52:34,740 --> 00:52:32,769
then then then week this is what we can

1216
00:52:39,269 --> 00:52:34,750
do we can we can those are the numbers

1217
00:52:41,009 --> 00:52:39,279
right there we excuse for identifying

1218
00:52:42,809 --> 00:52:41,019
functions because things that are in

1219
00:52:45,509 --> 00:52:42,819
like we don't know the function of any

1220
00:52:47,940 --> 00:52:45,519
color or in tuberculosis for example in

1221
00:52:49,680 --> 00:52:47,950
this case and you can look at it and

1222
00:52:51,930 --> 00:52:49,690
look and see what it interacts it and we

1223
00:52:53,759 --> 00:52:51,940
can predict interact with that and that

1224
00:52:56,099 --> 00:52:53,769
it may be does the same kind of function

1225
00:52:57,839 --> 00:52:56,109
and you can predict what I essential for

1226
00:53:00,990 --> 00:52:57,849
the organism so if you want to design a

1227
00:53:03,359 --> 00:53:01,000
new organism that that is useful in some

1228
00:53:06,120 --> 00:53:03,369
other environment we can we can we can

1229
00:53:07,799 --> 00:53:06,130
we can do that and we can also predict

1230
00:53:11,400 --> 00:53:07,809
first position attractions that's again

1231
00:53:14,880 --> 00:53:11,410
in microbiology issue like I said blue

1232
00:53:16,440 --> 00:53:14,890
benchmark colors so what are we doing I

1233
00:53:18,150 --> 00:53:16,450
mean they're combining all of this data

1234
00:53:21,059 --> 00:53:18,160
we combine the individual structure

1235
00:53:24,779 --> 00:53:21,069
structure function interaction leader

1236
00:53:26,999 --> 00:53:24,789
with the genome by data the gene array

1237
00:53:29,009 --> 00:53:27,009
and the functional genomics data and

1238
00:53:31,109 --> 00:53:29,019

here's a simple example of what happens

1239
00:53:33,390 --> 00:53:31,119
this is the lac operon working in tracks

1240
00:53:36,240 --> 00:53:33,400
the transcription factor it binds to the

1241
00:53:38,099 --> 00:53:36,250
DNA it coats the thing cause the mRNA

1242
00:53:41,579 --> 00:53:38,109
that interacts the protein and the

1243
00:53:44,730 --> 00:53:41,589
feedback link and in essential in

1244
00:53:47,190 --> 00:53:44,740
essence you as an organism and me as

1245
00:53:50,130 --> 00:53:47,200
organs are one big large feedback loops

1246
00:53:51,630 --> 00:53:50,140
so so there's that that's what we're

1247
00:53:54,650 --> 00:53:51,640
trying to replicate and you're

1248
00:53:59,009 --> 00:53:54,660
integrating a lot of data to do this and

1249
00:54:00,960 --> 00:53:59,019
some gang time one of the things that we

1250
00:54:06,749 --> 00:54:00,970
do believe in is making all our

1251
00:54:08,430 --> 00:54:06,759
algorithms and our work public at least

1252
00:54:12,509 --> 00:54:08,440
available on the web so other biologists

1253
00:54:15,720 --> 00:54:12,519
can use them and my kirkin was sitting

1254
00:54:17,579 --> 00:54:15,730
there has created a nice databases

1255
00:54:20,099 --> 00:54:17,589
actually create a web the web Trent end

1256
00:54:21,930 --> 00:54:20,109
of it is this is this is google for by

1257
00:54:23,999 --> 00:54:21,940
informatics and this was before Google

1258
00:54:27,990 --> 00:54:24,009
came up with Gmail and stuff like that

1259
00:54:30,630 --> 00:54:28,000
so Mike was ahead of the curve and and

1260
00:54:33,539 --> 00:54:30,640
and so he's gotten all this data in

1261
00:54:35,220 --> 00:54:33,549
there and these are the URLs for me this

1262
00:54:38,460 --> 00:54:35,230
is you all for all day that we've

1263
00:54:40,710 --> 00:54:38,470

analyzed 54 podium so far I think it's

1264
00:54:43,320 --> 00:54:40,720
now 65 in Detroit night

1265
00:54:44,670 --> 00:54:43,330
yes because the face of war and then the

1266
00:54:45,900 --> 00:54:44,680
individual if you want to hear every

1267
00:54:48,420 --> 00:54:45,910
single protein and you want to look at

1268
00:54:51,349 --> 00:54:48,430
it you can go to these servers and they

1269
00:54:53,910 --> 00:54:51,359
will predict the functional structure

1270
00:54:55,500 --> 00:54:53,920
and they give you a lot of detail about

1271
00:54:57,120 --> 00:54:55,510
all these things you can actually

1272
00:54:59,790 --> 00:54:57,130
visualize these interaction graphs using

1273
00:55:01,800 --> 00:54:59,800
a web browser again the whole goal is to

1274
00:55:04,589 --> 00:55:01,810
make everything life life easy for

1275
00:55:07,140 --> 00:55:04,599
biologists I'm going to go a little

1276
00:55:08,880 --> 00:55:07,150
overtime here but I want to talk about

1277
00:55:10,770 --> 00:55:08,890
some of the applications of what we're

1278
00:55:13,349 --> 00:55:10,780
doing so we've done some drug discovery

1279
00:55:16,260 --> 00:55:13,359
work so we predicted molecules that bind

1280
00:55:19,380 --> 00:55:16,270
to herpes prettiest there is no known

1281
00:55:21,359 --> 00:55:19,390
herpes protease inhibitor all herpes

1282
00:55:23,820 --> 00:55:21,369
drugs are what you call nucleoside

1283
00:55:28,859 --> 00:55:23,830
analogs they are like to use an analogy

1284
00:55:33,300 --> 00:55:28,869
like the like HIV reverse transcriptase

1285
00:55:36,630 --> 00:55:33,310
inhibitors that is like HIV like it's

1286
00:55:38,849 --> 00:55:36,640
the first HIV try to forget but but then

1287
00:55:41,580 --> 00:55:38,859
you combine them and you form a cocktail

1288
00:55:43,710 --> 00:55:41,590

so I'm sure you're familiar the concept

1289
00:55:46,829 --> 00:55:43,720
of the HIV cocktail and that's what

1290
00:55:50,160 --> 00:55:46,839
works against us so so we we have the

1291
00:55:52,020 --> 00:55:50,170
first herpes inhibitor that protease

1292
00:55:56,460 --> 00:55:52,030
inhibitor that would be combined with

1293
00:55:57,390 --> 00:55:56,470
the with the current herpes drugs that

1294
00:56:00,420 --> 00:55:57,400
could be combined with the current

1295
00:56:02,370 --> 00:56:00,430
herpes drugs and uses a cocktail and I

1296
00:56:04,020 --> 00:56:02,380
have a lot of external data so that was

1297
00:56:07,109 --> 00:56:04,030
a prediction but what you saw pictures

1298
00:56:09,930 --> 00:56:07,119
of that was the few predictions I have

1299
00:56:12,960 --> 00:56:09,940
lolok small data to show that it

1300
00:56:15,060 --> 00:56:12,970
inhibits HIV replication all there are

1301
00:56:19,109 --> 00:56:15,070
eight human herpesviruses believe it or

1302
00:56:20,250 --> 00:56:19,119
not chicken pox is is a how to slice if

1303
00:56:23,099 --> 00:56:20,260
you didn't know that and so

1304
00:56:25,050 --> 00:56:23,109
varicella-zoster cytomegalovirus is

1305
00:56:29,510 --> 00:56:25,060
about ten percent of transplant patients

1306
00:56:32,970 --> 00:56:29,520
kaposi sarcoma herpes virus as a

1307
00:56:34,890 --> 00:56:32,980
associated her piecewise is something

1308
00:56:36,810 --> 00:56:34,900
that most people infected with HIV a

1309
00:56:40,740 --> 00:56:36,820
coin factory with all these herpes

1310
00:56:42,270 --> 00:56:40,750
viruses and we can show that our drugs

1311
00:56:44,180 --> 00:56:42,280
works better than the current current

1312
00:56:46,410 --> 00:56:44,190
drugs are comparable to and

1313
00:56:48,570 --> 00:56:46,420

synergistically signature skiing achieve

1314
00:56:50,490 --> 00:56:48,580
this the synergy plot show that it

1315
00:56:53,120 --> 00:56:50,500
pretty much right here this is the one

1316
00:56:57,140 --> 00:56:53,130
where you combine our drug

1317
00:57:01,370 --> 00:56:57,150
with with with acyclovir the current

1318
00:57:04,759 --> 00:57:01,380
standard accepted drug while traxxas is

1319
00:57:07,730 --> 00:57:04,769
essentially a new patent form of

1320
00:57:10,039 --> 00:57:07,740
acyclovir if you want to call it that it

1321
00:57:11,390 --> 00:57:10,049
essentially describes all wise in a cell

1322
00:57:14,329 --> 00:57:11,400
cultures so these are salt culture

1323
00:57:16,670 --> 00:57:14,339
studies so we now killing a lot of mice

1324
00:57:18,890 --> 00:57:16,680
I mean that is infecting a lot of nice

1325
00:57:21,440 --> 00:57:18,900
with herpes which actually kills them

1326
00:57:23,450 --> 00:57:21,450
and then and then seeing what happens

1327
00:57:26,870 --> 00:57:23,460
with that so hopefully it'll work in

1328
00:57:28,759 --> 00:57:26,880
Mike's right now we also do this with

1329
00:57:31,370 --> 00:57:28,769
existing drugs that is we take existing

1330
00:57:34,430 --> 00:57:31,380
drugs again the reason for doing that is

1331
00:57:36,380 --> 00:57:34,440
because evolutions reused substrates

1332
00:57:40,460 --> 00:57:36,390
again and again and again so existing

1333
00:57:42,499 --> 00:57:40,470
drugs at work work better so we also all

1334
00:57:45,440 --> 00:57:42,509
have the pharmacology of data the toxic

1335
00:57:47,779 --> 00:57:45,450
data and so on and so I'll use that so

1336
00:57:51,230 --> 00:57:47,789
what does this mean why why am i showing

1337
00:57:55,430 --> 00:57:51,240
you this how what I I mean it's him to

1338
00:57:57,200 --> 00:57:55,440

say for drug discovery point I was

1339
00:58:02,329 --> 00:57:57,210
recently more than a grant to regenerate

1340
00:58:04,670 --> 00:58:02,339
the truth tooth and what we can do is

1341
00:58:08,120 --> 00:58:04,680
our techniques are so general and so

1342
00:58:11,029 --> 00:58:08,130
broad that we don't have to worry about

1343
00:58:12,769 --> 00:58:11,039
inhibition of a particular protein we

1344
00:58:14,960 --> 00:58:12,779
can induce a particular protein to do

1345
00:58:17,210 --> 00:58:14,970
something and we actually have a set of

1346
00:58:19,670 --> 00:58:17,220
compounds when people talk about the RNA

1347
00:58:21,620 --> 00:58:19,680
world or the DNA world or whatever

1348
00:58:24,170 --> 00:58:21,630
whatever hypothesis that you believe in

1349
00:58:26,240 --> 00:58:24,180
that originated in life it didn't it

1350
00:58:27,890 --> 00:58:26,250
didn't happen like that there were the

1351
00:58:29,390 --> 00:58:27,900
small molecules in the world in fact

1352
00:58:31,519 --> 00:58:29,400
there are probably a lot of other small

1353
00:58:33,140 --> 00:58:31,529
molecules in one and you need all these

1354
00:58:35,240 --> 00:58:33,150
small molecules induce the gene

1355
00:58:37,700 --> 00:58:35,250
expression signature that you want to

1356
00:58:39,890 --> 00:58:37,710
make the organism survive and we can do

1357
00:58:41,900 --> 00:58:39,900
that we can use a particular gene

1358
00:58:43,460 --> 00:58:41,910
expression signature using these small

1359
00:58:46,249 --> 00:58:43,470
molecule techniques docking techniques

1360
00:58:52,910 --> 00:58:46,259
that we developed I hope that point is

1361
00:58:56,089 --> 00:58:52,920
very clear to everyone I mean we yeah I

1362
00:58:58,970 --> 00:58:56,099
mean it David other molecules it wasn't

1363
00:59:00,499 --> 00:58:58,980

just RNA and DNA and proteins or what

1364
00:59:02,150 --> 00:59:00,509
came first or what came later they were

1365
00:59:04,069 --> 00:59:02,160
there are other molecules lolla

1366
00:59:05,530 --> 00:59:04,079
biological substrates that made life on

1367
00:59:08,470 --> 00:59:05,540
Earth possible

1368
00:59:10,900 --> 00:59:08,480
and those are extremely essential if you

1369
00:59:13,440 --> 00:59:10,910
want to talk about astrobiology and

1370
00:59:17,320 --> 00:59:13,450
sitting life on other planets and so on

1371
00:59:19,450 --> 00:59:17,330
ok and then nanotechnology so here's

1372
00:59:21,160 --> 00:59:19,460
another case where this is even more

1373
00:59:23,470 --> 00:59:21,170
abstract and this is actually have

1374
00:59:27,160 --> 00:59:23,480
relevance to extreme environments and so

1375
00:59:29,530 --> 00:59:27,170
on where we I mean so what I said to you

1376
00:59:31,090 --> 00:59:29,540
before is our predictions that have been

1377
00:59:33,360 --> 00:59:31,100
completely verified by X women

1378
00:59:36,220 --> 00:59:33,370
completely match with what we see

1379
00:59:38,200 --> 00:59:36,230
experimentally so it's one thing to talk

1380
00:59:39,940 --> 00:59:38,210
about computation logging and you know

1381
00:59:42,070 --> 00:59:39,950
publish a lot of papers on computational

1382
00:59:45,010 --> 00:59:42,080
modeling but if you don't get it

1383
00:59:47,410 --> 00:59:45,020
verified by X mental by observation then

1384
00:59:49,390 --> 00:59:47,420
then it's meaningless as far as I'm

1385
00:59:51,970 --> 00:59:49,400
concerned that's why Cass was initiated

1386
00:59:54,340 --> 00:59:51,980
and so the herpes stuff that I showed

1387
00:59:56,320 --> 00:59:54,350
you was an example we've done this for

1388
00:59:59,560 --> 00:59:56,330

malaria for dengue and we can show that

1389
01:00:01,840 --> 00:59:59,570
X parently that that these these results

1390
01:00:04,060 --> 01:00:01,850
that the prediction that we make are not

1391
01:00:06,400 --> 01:00:04,070
completely accurate but still failing

1392
01:00:08,560 --> 01:00:06,410
the top top bridge again this is a case

1393
01:00:12,010 --> 01:00:08,570
where predict a small peptides of

1394
01:00:15,010 --> 01:00:12,020
proteins a pint in organic substrates so

1395
01:00:16,930 --> 01:00:15,020
we are a carbonyl eyes baseball we are

1396
01:00:19,720 --> 01:00:16,940
we are carbon-based life-forms

1397
01:00:22,930 --> 01:00:19,730
carbon-based but there could be

1398
01:00:25,720 --> 01:00:22,940
silica-based got life forms factor con

1399
01:00:27,970 --> 01:00:25,730
computers but but there could be other

1400
01:00:29,620 --> 01:00:27,980
other other other life forms that that

1401
01:00:32,980 --> 01:00:29,630
you might think of that are not

1402
01:00:34,930 --> 01:00:32,990
carbon-based and we can design enzymes

1403
01:00:37,330 --> 01:00:34,940
and proteins to get around that and

1404
01:00:39,370 --> 01:00:37,340
we've done that in this particular case

1405
01:00:42,310 --> 01:00:39,380
the actually looking wards and design of

1406
01:00:43,810 --> 01:00:42,320
new proteins that bind tech parts that

1407
01:00:46,900 --> 01:00:43,820
don't have functions that have never

1408
01:00:49,650 --> 01:00:46,910
been observed in nature yet and we show

1409
01:00:52,240 --> 01:00:49,660
that piece buying the cords as predicted

1410
01:00:55,420 --> 01:00:52,250
the best one that is discovered expertly

1411
01:00:59,680 --> 01:00:55,430
is in black right here this is by law

1412
01:01:02,470 --> 01:00:59,690
exposed techniques a lot of work and we

1413
01:01:04,720 --> 01:01:02,480

just do our simulations we use use the

1414
01:01:06,460 --> 01:01:04,730
Explorer beta I mean I have to be honest

1415
01:01:08,650 --> 01:01:06,470
with you I music-related starting point

1416
01:01:11,920 --> 01:01:08,660
and then the you do our simulations and

1417
01:01:14,980 --> 01:01:11,930
our strongest binders this is again an

1418
01:01:17,410 --> 01:01:14,990
external result

1419
01:01:19,000 --> 01:01:17,420
an spr result I won't go into detail on

1420
01:01:21,370 --> 01:01:19,010
that but what you need to do this look

1421
01:01:23,380 --> 01:01:21,380
at is discover us our first predicted

1422
01:01:28,210 --> 01:01:23,390
binder called spine directs the car

1423
01:01:30,220 --> 01:01:28,220
binder is happens to be the strongest

1424
01:01:32,080 --> 01:01:30,230
one as we predicted we also need a

1425
01:01:34,180 --> 01:01:32,090
negative control way we took what we

1426
01:01:36,520 --> 01:01:34,190
thought were the weakest binders and we

1427
01:01:40,510 --> 01:01:36,530
show that there's a clear separation so

1428
01:01:44,020 --> 01:01:40,520
we might be off on s4 and s5 right here

1429
01:01:45,609 --> 01:01:44,030
but so you know it's it's still

1430
01:01:47,740 --> 01:01:45,619
inconsistent I mean there's a very clear

1431
01:01:49,030 --> 01:01:47,750
separation in fact I would say this is

1432
01:01:53,680 --> 01:01:49,040
one hundred percent agreement with

1433
01:01:56,170 --> 01:01:53,690
experiment or for predictions okay so

1434
01:01:57,850 --> 01:01:56,180
what is the feature I'm a little bit

1435
01:02:02,490 --> 01:01:57,860
over time but I since I started a little

1436
01:02:05,740 --> 01:02:02,500
late I can think a little birdie of that

1437
01:02:11,680 --> 01:02:05,750
I'll finish in a minute so what is the

1438
01:02:13,180 --> 01:02:11,690

feature the future is that the future is

1439
01:02:15,190 --> 01:02:13,190
that we have a lot of structural data

1440
01:02:18,220 --> 01:02:15,200
coming out I really believe in the

1441
01:02:20,820 --> 01:02:18,230
concept of atomic level modeling so it's

1442
01:02:23,740 --> 01:02:20,830
nice online mechanics at this point so

1443
01:02:25,210 --> 01:02:23,750
but I believe that's enough we don't

1444
01:02:26,920 --> 01:02:25,220
even need to call it quantum physics for

1445
01:02:28,960 --> 01:02:26,930
trying to understand how proteins work

1446
01:02:31,780 --> 01:02:28,970
but but that's that's just my prejudice

1447
01:02:34,570 --> 01:02:31,790
but I could be corrected and I might be

1448
01:02:36,550 --> 01:02:34,580
wrong but I think that proteins be here

1449
01:02:38,410 --> 01:02:36,560
now on a Newtonian level and we can

1450
01:02:40,420 --> 01:02:38,420
model them like that so this huge amount

1451
01:02:42,580 --> 01:02:40,430
of atomic level data coming out and we

1452
01:02:44,380 --> 01:02:42,590
can exploit the data there's a huge

1453
01:02:45,630 --> 01:02:44,390
amount of functional data exploring data

1454
01:02:47,710 --> 01:02:45,640
it's coming out we can exploit that

1455
01:02:51,070 --> 01:02:47,720
integrate that into our simulation

1456
01:02:54,130 --> 01:02:51,080
methods but the modern data that's being

1457
01:02:56,290 --> 01:02:54,140
pretty is so large that there is no

1458
01:02:58,150 --> 01:02:56,300
human brain in this world that can

1459
01:03:01,210 --> 01:02:58,160
process all of this you know the first

1460
01:03:03,550 --> 01:03:01,220
cast that was in 1994 where everyone did

1461
01:03:05,440 --> 01:03:03,560
badly in 1996 and linking that here

1462
01:03:08,080 --> 01:03:05,450
there was a human person who did better

1463
01:03:10,210 --> 01:03:08,090

than all the computers and that was

1464
01:03:12,849 --> 01:03:10,220
something that people value people were

1465
01:03:14,260 --> 01:03:12,859
no priding about it that there is a

1466
01:03:15,940 --> 01:03:14,270
human who can do better than all the

1467
01:03:19,090 --> 01:03:15,950
computers well that didn't last very

1468
01:03:19,760 --> 01:03:19,100
long because the computers got faster

1469
01:03:25,520 --> 01:03:19,770
and

1470
01:03:26,960 --> 01:03:25,530
programmers wrote better programs so how

1471
01:03:28,910 --> 01:03:26,970
do you man when you're talking about

1472
01:03:30,830 --> 01:03:28,920
more much more complex levels of

1473
01:03:32,990 --> 01:03:30,840
information how do you integrate all of

1474
01:03:34,940 --> 01:03:33,000
this and how do you give a semantic

1475
01:03:37,370 --> 01:03:34,950
minix and I think computers are the only

1476
01:03:39,080 --> 01:03:37,380
answer even now when you have used

1477
01:03:41,660 --> 01:03:39,090
Google you just have to type in a string

1478
01:03:43,790 --> 01:03:41,670
and you get some results you can't ask

1479
01:03:46,160 --> 01:03:43,800
Google a question and get a result for

1480
01:03:48,620 --> 01:03:46,170
it it doesn't give you a semantic

1481
01:03:50,120 --> 01:03:48,630
meaning it only gives you a set of

1482
01:03:51,590 --> 01:03:50,130
results and what we are trying to

1483
01:03:53,570 --> 01:03:51,600
produce in the bio verse that Mike is

1484
01:03:55,970 --> 01:03:53,580
working on is to produce the biological

1485
01:03:59,300 --> 01:03:55,980
model and when you when we can do that

1486
01:04:01,070 --> 01:03:59,310
going back to the astrobiology aspects

1487
01:04:04,010 --> 01:04:01,080
of it when we can do that we can

1488
01:04:05,660 --> 01:04:04,020

engineer new organisms we can engineer

1489
01:04:07,220 --> 01:04:05,670
new organized for other plants we cannot

1490
01:04:09,920 --> 01:04:07,230
engineer organism any any any

1491
01:04:11,510 --> 01:04:09,930
environment I mean this is this is still

1492
01:04:13,940 --> 01:04:11,520
a long way away but we're that's that's

1493
01:04:15,140 --> 01:04:13,950
what I researchers directly towards a

1494
01:04:19,430 --> 01:04:15,150
loose or all the tools that we

1495
01:04:21,560 --> 01:04:19,440
developing so I so so my fundamental

1496
01:04:24,020 --> 01:04:21,570
message is a long morning they won't you

1497
01:04:26,810 --> 01:04:24,030
guys take home is that morning proteins

1498
01:04:28,700 --> 01:04:26,820
and proteome structure and function at

1499
01:04:31,130 --> 01:04:28,710
the atomic level at really at atomic

1500
01:04:32,780 --> 01:04:31,140
level the Newtonian level is necessary

1501
01:04:35,330 --> 01:04:32,790
to understand the relation is routine

1502
01:04:37,130 --> 01:04:35,340
you know single molecules single

1503
01:04:41,690 --> 01:04:37,140
functional molecules systems pathways

1504
01:04:44,750 --> 01:04:41,700
cells and emily organisms and like i

1505
01:04:46,850 --> 01:04:44,760
said this is an older talk now would say

1506
01:04:49,270 --> 01:04:46,860
more than four teens and modeling DNA

1507
01:04:51,800 --> 01:04:49,280
and modeling RNA and small molecules i

1508
01:04:57,470 --> 01:04:51,810
want to acknowledge all the people in my

1509
01:04:59,180 --> 01:04:57,480
group and bunch of them and and this

1510
01:05:01,970 --> 01:04:59,190
again are all the slides the

1511
01:05:09,999 --> 01:05:01,980
collaborators and also my funding

1512
01:05:21,160 --> 01:05:16,880
the question Phillies cap use these

1513
01:05:26,809 --> 01:05:21,170

methods the moral protein folding weird

1514
01:05:29,029 --> 01:05:26,819
yeah absolutely well you can the answer

1515
01:05:32,749 --> 01:05:29,039
is you can whether you'll get it right

1516
01:05:35,480 --> 01:05:32,759
or not is another issue that I mean may

1517
01:05:39,109 --> 01:05:35,490
I you know it yeah you can do it and

1518
01:05:40,849 --> 01:05:39,119
that's one of the points of trying to

1519
01:05:43,099 --> 01:05:40,859
identify what are the important rescues

1520
01:05:44,870 --> 01:05:43,109
and how would these SGS behave under

1521
01:05:47,240 --> 01:05:44,880
different conditions I think that you

1522
01:05:50,390 --> 01:05:47,250
can you can use a combination of all

1523
01:05:52,400 --> 01:05:50,400
this information to do it right now it's

1524
01:05:54,650 --> 01:05:52,410
probably not automatic probably has to

1525
01:05:57,230 --> 01:05:54,660
be done manually and Aaron is again

1526
01:05:59,299 --> 01:05:57,240
looking at it very in very close detail

1527
01:06:00,980 --> 01:05:59,309
but it can be done and that's the whole

1528
01:06:04,549 --> 01:06:00,990
idea I mean that's where we're going

1529
01:06:05,990 --> 01:06:04,559
towards and I can tell you that John

1530
01:06:07,670 --> 01:06:06,000
layers work with me and we are trying to

1531
01:06:10,339 --> 01:06:07,680
model some of these proteins and it's

1532
01:06:13,339 --> 01:06:10,349
been really hard for us because he works

1533
01:06:15,620 --> 01:06:13,349
with extreme extreme puddings that work

1534
01:06:18,349 --> 01:06:15,630
in extreme environments and we don't

1535
01:06:21,230 --> 01:06:18,359
have a lot of data on that and so a lot

1536
01:06:24,230 --> 01:06:21,240
of our approaches knowledge base so I

1537
01:06:26,059 --> 01:06:24,240
mean it's a long drawn-out answer but as

1538
01:06:27,769 --> 01:06:26,069

we get more and more data as John

1539
01:06:29,509 --> 01:06:27,779
produces more data we can we can

1540
01:06:31,849 --> 01:06:29,519
incorporate the data into a simulation

1541
01:06:33,680 --> 01:06:31,859
protocols so it's a collaboration

1542
01:06:40,549 --> 01:06:33,690
between experimental and computational

1543
01:06:44,029 --> 01:06:40,559
values that's an iterative process here

1544
01:06:45,970 --> 01:06:44,039
term future for understanding a function

1545
01:06:51,049 --> 01:06:45,980
of the four different people forever

1546
01:06:52,789 --> 01:06:51,059
alone being trained at the protein I

1547
01:06:54,440 --> 01:06:52,799
think we can understand the function I

1548
01:06:56,690 --> 01:06:54,450
mean again what do you mean by

1549
01:06:58,460 --> 01:06:56,700
understanding function I mean so this

1550
01:06:59,930 --> 01:06:58,470
this goes into why we developed a

1551
01:07:01,849 --> 01:06:59,940
functional signature pressure the

1552
01:07:06,200 --> 01:07:01,859
functional signature is what kind i

1553
01:07:09,109 --> 01:07:06,210
would say bye-bye us by real so you talk

1554
01:07:11,299 --> 01:07:09,119
overly immunoglobulins what did they

1555
01:07:13,099 --> 01:07:11,309
bind to antigens but what did they all

1556
01:07:15,140 --> 01:07:13,109
say do they are part of the different

1557
01:07:17,029 --> 01:07:15,150
system what do you call it what do you

1558
01:07:19,370 --> 01:07:17,039
call that function as english is not

1559
01:07:20,440 --> 01:07:19,380
enough to describe the function of

1560
01:07:23,020 --> 01:07:20,450
protein

1561
01:07:25,030 --> 01:07:23,030
so it needs to be mathematical and i

1562
01:07:27,490 --> 01:07:25,040
would say that we can actually model the

1563
01:07:30,400 --> 01:07:27,500

function of about more than fifty

1564
01:07:33,430 --> 01:07:30,410
percent of a given podium right now and

1565
01:07:36,520 --> 01:07:33,440
as again as more xml data is available

1566
01:07:39,040 --> 01:07:36,530
we can keep incorporating that into our

1567
01:07:40,900 --> 01:07:39,050
simulations and but that's we need

1568
01:07:43,660 --> 01:07:40,910
explanation there's there's no doubt

1569
01:07:46,480 --> 01:07:43,670
about that and our methods that there

1570
01:07:48,339 --> 01:07:46,490
was to handle that but function the word

1571
01:07:50,920 --> 01:07:48,349
function is is is actually very

1572
01:07:52,480 --> 01:07:50,930
arbitrary and if you go by English

1573
01:07:54,579 --> 01:07:52,490
language definitions i think is actually

1574
01:07:59,500 --> 01:07:54,589
a wrong what we need quantitative

1575
01:08:02,250 --> 01:07:59,510
definitions of functions also driving

1576
01:08:07,000 --> 01:08:02,260
and you have a a finite number of

1577
01:08:09,790 --> 01:08:07,010
security structural options absolutely

1578
01:08:13,839 --> 01:08:09,800
that they do I have why are we not been

1579
01:08:17,050 --> 01:08:13,849
able to actually protect I hate you for

1580
01:08:20,200 --> 01:08:17,060
an function but we have application of

1581
01:08:23,079 --> 01:08:20,210
that we're protein or living organ but

1582
01:08:24,789 --> 01:08:23,089
we have and that's that's that's what

1583
01:08:26,410 --> 01:08:24,799
that's that's what happens when you take

1584
01:08:29,079 --> 01:08:26,420
the evolutionary history of the organism

1585
01:08:31,599 --> 01:08:29,089
into account that when you to take the

1586
01:08:33,880 --> 01:08:31,609
fact that the that functional important

1587
01:08:35,470 --> 01:08:33,890
amino acids are functionally frustrated

1588
01:08:37,539 --> 01:08:35,480

that that is structurally frustrated

1589
01:08:39,729 --> 01:08:37,549
that is there they're not stable until

1590
01:08:41,590 --> 01:08:39,739
they're in the functional form and our

1591
01:08:43,180 --> 01:08:41,600
accuracy improves a lot and it's

1592
01:08:46,950 --> 01:08:43,190
probably a few other factors that we are

1593
01:08:51,459 --> 01:08:46,960
missing that we don't know yet and again

1594
01:08:54,190 --> 01:08:51,469
thought of Alan's thesis so yeah but but

1595
01:08:57,249 --> 01:08:54,200
that's but we have we are probably among

1596
01:09:00,090 --> 01:08:57,259
the first people do that where we are

1597
01:09:01,930 --> 01:09:00,100
able to take things that look the same

1598
01:09:07,709 --> 01:09:01,940
you would think they'd do the same

1599
01:09:13,860 --> 01:09:10,360
looking at your network of the proteome

1600
01:09:19,030 --> 01:09:17,290
part of thinking about think about the

1601
01:09:22,690 --> 01:09:19,040
human when it's it's about a three

1602
01:09:24,730 --> 01:09:22,700
million about crayons our prion protein

1603
01:09:27,040 --> 01:09:24,740
can actually affect a structural change

1604
01:09:28,180 --> 01:09:27,050
in southern front would you like to

1605
01:09:36,430 --> 01:09:28,190
discuss that from an evolutionary

1606
01:09:39,130 --> 01:09:36,440
perspective well so you know so we in a

1607
01:09:42,099 --> 01:09:39,140
sense where cherry picking you know we I

1608
01:09:43,900 --> 01:09:42,109
am good at doing that let's let's put it

1609
01:09:46,900 --> 01:09:43,910
that way where we are taking the

1610
01:09:50,530 --> 01:09:46,910
low-hanging fruit so do you I mean

1611
01:09:53,320 --> 01:09:50,540
infamy so you're you know in a sense all

1612
01:09:55,420 --> 01:09:53,330
of this is information transfer right so

1613
01:09:57,640 --> 01:09:55,430

we go to a channel to be and things like

1614
01:09:59,200 --> 01:09:57,650
that everything's about information

1615
01:10:00,790 --> 01:09:59,210
transfer so we're transferring

1616
01:10:05,260 --> 01:10:00,800
information from ourselves to our

1617
01:10:06,940 --> 01:10:05,270
progeny and so on and prions well

1618
01:10:09,730 --> 01:10:06,950
Stanley president won a nobel prize for

1619
01:10:13,450 --> 01:10:09,740
that is a way of transferring

1620
01:10:16,080 --> 01:10:13,460
information and can we model it actually

1621
01:10:18,550 --> 01:10:16,090
we can we can more to let that process

1622
01:10:20,950 --> 01:10:18,560
fairly well that's a well-established

1623
01:10:25,540 --> 01:10:20,960
process but again it's one of those

1624
01:10:27,520 --> 01:10:25,550
cases where there's not nothing else

1625
01:10:29,560 --> 01:10:27,530
besides prions like that you know so

1626
01:10:32,980 --> 01:10:29,570
it's it's it's it could be an over

1627
01:10:36,040 --> 01:10:32,990
testing kiss you know mean or or

1628
01:10:37,960 --> 01:10:36,050
training of training case so we can we

1629
01:10:39,460 --> 01:10:37,970
can I can take up you know prion protein

1630
01:10:41,500 --> 01:10:39,470
which adopts to conformational States

1631
01:10:43,600 --> 01:10:41,510
and I can make it out up into one

1632
01:10:45,670 --> 01:10:43,610
controversy I can see how it induces are

1633
01:10:48,240 --> 01:10:45,680
the protein conformational changes and i

1634
01:10:50,500 --> 01:10:48,250
can predict all of that installation but

1635
01:10:52,420 --> 01:10:50,510
then i have the right answer in front of

1636
01:10:56,380 --> 01:10:52,430
me but maybe i wrote the algorithm to

1637
01:11:00,930 --> 01:10:56,390
make their answer you know how do you be

1638
01:11:07,420 --> 01:11:04,780

yeah now being babying i think that that

1639
01:11:08,920 --> 01:11:07,430
I mean of means since since such a such

1640
01:11:10,660 --> 01:11:08,930
a general audience i think that is that

1641
01:11:12,610 --> 01:11:10,670
is really the main issue i think it's

1642
01:11:16,150 --> 01:11:12,620
really important to be very very very

1643
01:11:18,220 --> 01:11:16,160
hard son self-critical i think we we

1644
01:11:20,980 --> 01:11:18,230
suffer from that in the computation

1645
01:11:22,870 --> 01:11:20,990
field you know you and now what wasn't

1646
01:11:24,730 --> 01:11:22,880
it Robert Milligan did the oil drop

1647
01:11:27,280 --> 01:11:24,740
explain just going back to physic class

1648
01:11:30,040 --> 01:11:27,290
moaning you know he'd drop data points

1649
01:11:31,390 --> 01:11:30,050
from his blood just to show that he was

1650
01:11:34,390 --> 01:11:31,400
right but he was right he's right about

1651
01:11:39,670 --> 01:11:34,400
a charger truck but I do not encourage