
AbSciCon
2019

1
00:00:00,790 --> 00:00:07,320
[Music]

2
00:00:12,510 --> 00:00:09,350
[Applause]

3
00:00:15,450 --> 00:00:12,520
and so I'm gonna be talking about random

4
00:00:18,060 --> 00:00:15,460
polypeptide sequences and their

5
00:00:22,200 --> 00:00:18,070
relevance for the origins of life and

6
00:00:25,589 --> 00:00:22,210
I'm just gonna start straight there

7
00:00:27,600 --> 00:00:25,599
without not much introduction so I would

8
00:00:30,330 --> 00:00:27,610
like to ask you to imagine first that

9
00:00:32,940 --> 00:00:30,340
the circle here includes all the

10
00:00:37,979 --> 00:00:32,950
biological sequences that our life uses

11
00:00:42,150 --> 00:00:37,989
and we know that these sequences occupy

12
00:00:44,910 --> 00:00:42,160
only a very tiny part of the possible

13
00:00:51,900 --> 00:00:44,920

sequence space and of course this is

14
00:00:54,049 --> 00:00:51,910
terribly out of out of the French but

15
00:00:56,790 --> 00:00:54,059
this is basically just to illustrate

16
00:00:58,860 --> 00:00:56,800
what we are interested in so we are

17
00:01:03,840 --> 00:00:58,870
interested and what actually lies

18
00:01:07,260 --> 00:01:03,850
outside this box or this circle at this

19
00:01:09,990 --> 00:01:07,270
case in the random sequence space and

20
00:01:12,330 --> 00:01:10,000
especially when it comes to what could

21
00:01:17,819 --> 00:01:12,340
have kind of been lying around doing the

22
00:01:20,370 --> 00:01:17,829
early origins so to start whatever

23
00:01:23,010 --> 00:01:20,380
exploration are we basically looked at

24
00:01:25,620 --> 00:01:23,020
we performed our kind of like a scarce

25
00:01:27,719 --> 00:01:25,630
experimental sampling looking around the

26
00:01:32,100 --> 00:01:27,729
biological space at some specific

27
00:01:36,120 --> 00:01:32,110
sequences that are out there and to

28
00:01:38,459 --> 00:01:36,130
start we generated 10,000 random

29
00:01:41,760 --> 00:01:38,469
sequences they were all of the same

30
00:01:45,240 --> 00:01:41,770
length 100 amino acids and the natural

31
00:01:48,899 --> 00:01:45,250
amino acid distribution and we performed

32
00:01:51,620 --> 00:01:48,909
secondary structure prediction on all of

33
00:01:54,810 --> 00:01:51,630
these using multiple predictors and

34
00:01:58,380 --> 00:01:54,820
based on these predictions we sorted the

35
00:02:02,719 --> 00:01:58,390
data out and selected some sequences

36
00:02:07,260 --> 00:02:02,729
that we looked at experimentally and

37
00:02:09,749 --> 00:02:07,270
these results showed the by informatique

38
00:02:12,090 --> 00:02:09,759

prediction of the secondary structure

39
00:02:15,720 --> 00:02:12,100
and the first panel is our random

40
00:02:17,970 --> 00:02:15,730
sequence data set and we used five

41
00:02:20,130 --> 00:02:17,980
different predictors and looked always

42
00:02:21,300 --> 00:02:20,140
at the alpha helical and beta sheet

43
00:02:23,309 --> 00:02:21,310
content

44
00:02:25,500 --> 00:02:23,319
and we did the same thing for some

45
00:02:29,040 --> 00:02:25,510
control groups the first were fragments

46
00:02:31,860 --> 00:02:29,050
from the PDP data set database then

47
00:02:35,570 --> 00:02:31,870
fragments from the uniprot database and

48
00:02:38,640 --> 00:02:35,580
also fragments from the database of

49
00:02:41,550 --> 00:02:38,650
disordered proteins and a short message

50
00:02:44,270 --> 00:02:41,560
here is that secondary structure

51
00:02:49,800 --> 00:02:44,280
actually seems to be quite abundant in

52
00:02:52,440 --> 00:02:49,810
random sequence base this plot here

53
00:02:54,930 --> 00:02:52,450
shows the same random data set in a

54
00:03:00,809 --> 00:02:54,940
slightly different way so on the y-axis

55
00:03:03,630 --> 00:03:00,819
here we have secondary structure content

56
00:03:07,350 --> 00:03:03,640
and on the x-axis we have the predicted

57
00:03:10,650 --> 00:03:07,360
disorder of these sequences so basically

58
00:03:14,370 --> 00:03:10,660
we see a whole range of sequences in our

59
00:03:16,259 --> 00:03:14,380
random data set and we have some

60
00:03:17,970 --> 00:03:16,269
sequences with high secondary structure

61
00:03:21,210 --> 00:03:17,980
content and sound with low secondary

62
00:03:23,280 --> 00:03:21,220
structure content and we selected 45

63
00:03:27,930 --> 00:03:23,290

sequences here those are the ones and in

64
00:03:31,020 --> 00:03:27,940
color for some experiments ensured this

65
00:03:34,080 --> 00:03:31,030
blood here summarizes all these

66
00:03:36,120 --> 00:03:34,090
experiments so we looked at expression

67
00:03:40,289 --> 00:03:36,130
and solubility in e.coli

68
00:03:42,750 --> 00:03:40,299
and also secondary structure using CD on

69
00:03:44,309 --> 00:03:42,760
purified proteins and I don't really

70
00:03:49,410 --> 00:03:44,319
want to go into much detail there is a

71
00:03:51,599 --> 00:03:49,420
lot of detail there but to summarize we

72
00:03:54,000 --> 00:03:51,609
basically confirm the secondary

73
00:03:56,370 --> 00:03:54,010
structure is quite abundant and random

74
00:03:59,400 --> 00:03:56,380
sequences and we did not really expect

75
00:04:01,710 --> 00:03:59,410
to see that and also there are random

76
00:04:04,979 --> 00:04:01,720
sequences that actually hi have hi

77
00:04:07,140 --> 00:04:04,989
disorder at the content seem to be

78
00:04:11,099 --> 00:04:07,150
better tolerated by living cells they

79
00:04:12,900 --> 00:04:11,109
have low aggregation propensity and we

80
00:04:15,539 --> 00:04:12,910
hypothesized that these could actually

81
00:04:21,330 --> 00:04:15,549
make them better progenitors of syllable

82
00:04:25,710 --> 00:04:21,340
and functional proteins so basically

83
00:04:28,360 --> 00:04:25,720
based on this little search we concluded

84
00:04:34,000 --> 00:04:28,370
that it's worth looking out there

85
00:04:36,030 --> 00:04:34,010
and and particularly we decided we're

86
00:04:40,300 --> 00:04:36,040
gonna do this in a more systematic way

87
00:04:43,780 --> 00:04:40,310
we are interested and what we could find

88
00:04:46,330 --> 00:04:43,790

from different alphabets and especially

89
00:04:49,689 --> 00:04:46,340
those amino acid alphabets that could be

90
00:04:52,540 --> 00:04:49,699
more relevant during early origins doing

91
00:04:55,900 --> 00:04:52,550
evolution of proteins so basically we

92
00:04:58,870 --> 00:04:55,910
wanted to explore our different subsets

93
00:05:05,230 --> 00:04:58,880
of the after protein of the sequence

94
00:05:09,040 --> 00:05:05,240
base so early libraries are designed on

95
00:05:10,840 --> 00:05:09,050
a DNA level using degenerate codons and

96
00:05:14,830 --> 00:05:10,850
we started collaborating here with

97
00:05:17,379 --> 00:05:14,840
Sasuke Fukushima from LC and the

98
00:05:19,510 --> 00:05:17,389
degenerate codons basically control the

99
00:05:22,810 --> 00:05:19,520
amino acid composition of our libraries

100
00:05:27,909 --> 00:05:22,820
so we can design libraries from

101
00:05:31,529 --> 00:05:27,919
different subsets of amino acids and to

102
00:05:34,750 --> 00:05:31,539
do this as precisely as possible we

103
00:05:36,820 --> 00:05:34,760
developed an algorithm for this called

104
00:05:40,719 --> 00:05:36,830
the coot the degenerate codon

105
00:05:43,770 --> 00:05:40,729
optimization tool and this is roughly

106
00:05:49,710 --> 00:05:43,780
what it looks like so what we can do is

107
00:05:56,400 --> 00:05:49,720
basically find solutions to almost any

108
00:05:59,200 --> 00:05:56,410
alphabet and library lengths we can also

109
00:06:02,250 --> 00:05:59,210
optimize the codon usage to for

110
00:06:05,260 --> 00:06:02,260
expression at different organisms and

111
00:06:07,540 --> 00:06:05,270
also remove some codons for reassignment

112
00:06:10,300 --> 00:06:07,550
to be able to bring in some of the

113
00:06:13,180 --> 00:06:10,310

unnatural amino acids so amino acids

114
00:06:16,330 --> 00:06:13,190
that could be pre Baddeley very relevant

115
00:06:22,839 --> 00:06:16,340
but are not part of our genetic coding

116
00:06:26,650 --> 00:06:22,849
system at the moment so using this using

117
00:06:30,250 --> 00:06:26,660
this tool we basically started making

118
00:06:33,190 --> 00:06:30,260
some of these libraries and I guess one

119
00:06:35,920 --> 00:06:33,200
of the newest thing here is that we have

120
00:06:37,630 --> 00:06:35,930
been able to express enough quantity of

121
00:06:41,950 --> 00:06:37,640
these libraries using cell free

122
00:06:44,230 --> 00:06:41,960
expression system and purify this is a

123
00:06:47,140 --> 00:06:44,240
sample of one of these libraries and I

124
00:06:50,290 --> 00:06:47,150
know this looks very ugly but the reason

125
00:06:52,450 --> 00:06:50,300
is that we have a whole distribution of

126
00:06:54,790 --> 00:06:52,460
molecular weights within this within

127
00:06:57,070 --> 00:06:54,800
this band so that's basically what you

128
00:07:01,330 --> 00:06:57,080
see here on the model spectrum as what

129
00:07:03,730 --> 00:07:01,340
we have here in the sample and of course

130
00:07:07,089 --> 00:07:03,740
we are interested and libraries that

131
00:07:07,870 --> 00:07:07,099
would best mimic what was around kind of

132
00:07:11,620 --> 00:07:07,880
lying around

133
00:07:14,140 --> 00:07:11,630
during the early origins and while we do

134
00:07:17,050 --> 00:07:14,150
have some ideas what libraries we would

135
00:07:19,810 --> 00:07:17,060
like to make here I also wanted to use

136
00:07:21,580 --> 00:07:19,820
the chance of actually being here being

137
00:07:24,550 --> 00:07:21,590
able to present and thank you to the

138
00:07:28,810 --> 00:07:24,560

organizers for that and kind of invite

139
00:07:31,450 --> 00:07:28,820
you to now give us some some feedback on

140
00:07:33,399 --> 00:07:31,460
this because we are quite often ask like

141
00:07:36,129 --> 00:07:33,409
why these amino acids and why not the

142
00:07:39,370 --> 00:07:36,139
others and why this long and not that

143
00:07:41,469 --> 00:07:39,380
long so I would like to really invite

144
00:07:46,270 --> 00:07:41,479
you to come and talk to us at some point

145
00:07:50,080 --> 00:07:46,280
and tell us if you have ideas about what

146
00:07:53,320 --> 00:07:50,090
it's more relevant and before I do that

147
00:07:56,620 --> 00:07:53,330
I'd like to acknowledge especially my

148
00:07:59,230 --> 00:07:56,630
co-workers Koski Fukushima from Elsi and

149
00:08:01,719 --> 00:07:59,240
also Stephen Freese who is newly working

150
00:08:03,189 --> 00:08:01,729
with us on this project from John

151
00:08:06,399 --> 00:08:03,199
Hopkins University

152
00:08:09,939 --> 00:08:06,409
I would also acknowledge the people that

153
00:08:12,219 --> 00:08:09,949
I work with especially these guys from

154
00:08:15,219 --> 00:08:12,229
my group at the Charles University and

155
00:08:18,490 --> 00:08:15,229
that's in Prague and it's an ancient

156
00:08:21,939 --> 00:08:18,500
University but we work at this new

157
00:08:22,839 --> 00:08:21,949
campus close to Prague so thank you for

158
00:08:31,459 --> 00:08:22,849
your attention

159
00:09:09,319 --> 00:08:34,050
Thank You Clara we have a we have time

160
00:09:09,329 --> 00:09:26,800
yeah they were all 100 amino acid long

161
00:09:26,810 --> 00:09:36,309
[Music]

162
00:09:43,639 --> 00:09:40,729
yeah so the first study that I talked

163
00:09:46,129 --> 00:09:43,649

about was basically just looking at what

164
00:09:49,009 --> 00:09:46,139
we can find at the random secrets place

165
00:09:52,819 --> 00:09:49,019
but not that much relevance to origins

166
00:09:55,309 --> 00:09:52,829
of life so we chose that line of protein

167
00:09:58,699 --> 00:09:55,319
basically just because it's easier to

168
00:10:01,909 --> 00:09:58,709
break with and it was for us the first

169
00:10:03,590 --> 00:10:01,919
kind of little search so the proteins

170
00:10:07,909 --> 00:10:03,600
that we want to be looking at now would

171
00:10:10,879 --> 00:10:07,919
be much shorter yeah ranging from

172
00:10:20,100 --> 00:10:10,889
something between 20 and 60 is what we

173
00:10:20,110 --> 00:10:32,160
yeah

174
00:10:32,170 --> 00:10:35,810
hmm

175
00:10:35,820 --> 00:11:00,540
this sort of observe

176
00:11:05,199 --> 00:11:03,189
yes thank you this is something you have

177
00:11:08,019 --> 00:11:05,209
been thinking about a lot and talking to

178
00:11:10,269 --> 00:11:08,029
the disordered kind of protein community

179
00:11:12,129 --> 00:11:10,279
about it and I do agree that the

180
00:11:15,250 --> 00:11:12,139
disorder that we see is different than

181
00:11:18,879 --> 00:11:15,260
the disorder that they see first thing

182
00:11:21,910 --> 00:11:18,889
there is a very strong compositional

183
00:11:25,629 --> 00:11:21,920
bias in today's IDP proteins that are

184
00:11:28,389 --> 00:11:25,639
mostly eukaryotic and yeah of course

185
00:11:42,040 --> 00:11:28,399
it's so usually very functional what we

186
00:11:46,329 --> 00:11:42,050
see is a different case of disorder hi

187
00:11:51,069 --> 00:11:46,339
Mottola yes so we have so far been

188
00:11:54,069 --> 00:11:51,079

looking at comparing the full alphabet

189
00:11:56,470 --> 00:11:54,079
with different versions of what's

190
00:12:00,370 --> 00:11:56,480
considered the early alphabet meaning

191
00:12:04,480 --> 00:12:00,380
mostly just those amino acids there are

192
00:12:08,800 --> 00:12:04,490
biological today coded and were

193
00:12:10,870 --> 00:12:08,810
considered to be prebiotic available so

194
00:12:15,819 --> 00:12:10,880
that's what have we been doing so far

195
00:12:18,370 --> 00:12:15,829
but it's mostly work in progress right

196
00:12:21,069 --> 00:12:18,380
so I had a question um so it for your

197
00:12:23,769 --> 00:12:21,079
newer libraries do you have a way of

198
00:12:24,790 --> 00:12:23,779
dealing dealing with sort of single

199
00:12:26,530 --> 00:12:24,800
nucleotide insertions or deletions

200
00:12:28,300 --> 00:12:26,540
because I could imagine that if your if

201
00:12:29,949 --> 00:12:28,310
your library say when it's transcribed

202
00:12:31,840 --> 00:12:29,959
and your cell free translation system

203
00:12:34,720 --> 00:12:31,850
picks up a single insertion or deletion

204
00:12:35,829 --> 00:12:34,730
throws off your whole codon design is is

205
00:12:37,660 --> 00:12:35,839
there something about the system that

206
00:12:41,530 --> 00:12:37,670
would prevent that those from from

207
00:12:43,620 --> 00:12:41,540
coming up yes that's a good question it

208
00:12:47,650 --> 00:12:43,630
doesn't seem to be happening too much

209
00:12:51,069 --> 00:12:47,660
within our new library so we use we

210
00:12:53,879 --> 00:12:51,079
didn't use a lot of PCR not too much

211
00:12:57,759 --> 00:12:53,889
anyway and try to use high fidelity

212
00:13:01,300 --> 00:12:57,769
enzymes on the way so

213
00:13:03,220 --> 00:13:01,310

but after sequencing our library on the