
AbSciCon
2019

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

2
00:00:14,189 --> 00:00:09,199
[Applause]

3
00:00:15,499 --> 00:00:14,199
okay thank you so I'm as as I was

4
00:00:18,450 --> 00:00:15,509
introduced here to talk about the

5
00:00:21,839 --> 00:00:18,460
universal genetic code and the way I

6
00:00:23,700 --> 00:00:21,849
normally introduce this project to pure

7
00:00:25,589 --> 00:00:23,710
biochemists as I first have to convince

8
00:00:28,019 --> 00:00:25,599
them that this is something that might

9
00:00:29,730 --> 00:00:28,029
have evolved and did not magically

10
00:00:33,150 --> 00:00:29,740
appear overnight but I think I don't

11
00:00:36,180 --> 00:00:33,160
need to convince this room of that the

12
00:00:38,670 --> 00:00:36,190
very earliest life likely used the

13
00:00:41,789 --> 00:00:38,680

genetic code that was more simple than

14
00:00:43,859 --> 00:00:41,799
the universal code we know today likely

15
00:00:46,770 --> 00:00:43,869
comprising of just a handful of amino

16
00:00:49,380 --> 00:00:46,780
acids and it complexified over time

17
00:00:51,450 --> 00:00:49,390
adding more and more amino acids as it

18
00:00:54,390 --> 00:00:51,460
went there are a bunch of different

19
00:00:57,000 --> 00:00:54,400
theories for the order in which these

20
00:00:59,960 --> 00:00:57,010
amino acids were added these theories

21
00:01:03,899 --> 00:00:59,970
include the number of codons they have

22
00:01:06,450 --> 00:01:03,909
the GC richness of their codons the

23
00:01:07,950 --> 00:01:06,460
chemical simplicity of the amino acids

24
00:01:10,320 --> 00:01:07,960
themselves and whether they've been

25
00:01:13,590 --> 00:01:10,330
discovered on meteorites or synthesized

26
00:01:15,300 --> 00:01:13,600
in the lab finally we can look for clues

27
00:01:18,170 --> 00:01:15,310
and metabolisms so the order in which

28
00:01:21,840 --> 00:01:18,180
they can be synthesized in biology so

29
00:01:24,060 --> 00:01:21,850
there's an excess of 60 of these

30
00:01:27,690 --> 00:01:24,070
different hypotheses for the order of

31
00:01:30,510 --> 00:01:27,700
the codes evolution and for my work

32
00:01:34,020 --> 00:01:30,520
today we're utilizing a meta analysis

33
00:01:36,270 --> 00:01:34,030
performed by Trevor North in 2004 and he

34
00:01:39,120 --> 00:01:36,280
took all of these 16 amino acids and

35
00:01:41,880 --> 00:01:39,130
since synthesized them into a consensus

36
00:01:44,180 --> 00:01:41,890
order and so this year represents a

37
00:01:48,630 --> 00:01:44,190
single hypothesis for the codes

38
00:01:50,700 --> 00:01:48,640

evolution and today I'm going to be

39
00:01:52,770 --> 00:01:50,710
interrogating that and at the same time

40
00:01:56,760 --> 00:01:52,780
interrogating a more broad hypothesis

41
00:01:58,830 --> 00:01:56,770
can early genetic codes produce proteins

42
00:02:04,219 --> 00:01:58,840
that can support functions that are

43
00:02:06,980 --> 00:02:04,229
essential for life and to do that I took

44
00:02:10,190 --> 00:02:06,990
theoretical snapshots through this hyper

45
00:02:12,990 --> 00:02:10,200
hypothesized consensus order and

46
00:02:15,920 --> 00:02:13,000
generated libraries of randomized

47
00:02:18,780 --> 00:02:15,930
proteins that represent these different

48
00:02:21,690 --> 00:02:18,790
code hypothetical codes throughout the

49
00:02:24,360 --> 00:02:21,700
evolution so my most simple

50
00:02:27,660 --> 00:02:24,370
code has only the five most ancient

51
00:02:31,350 --> 00:02:27,670
amino acids and then the nine and then

52
00:02:34,339 --> 00:02:31,360
the sixteen lakhs just those final four

53
00:02:37,440 --> 00:02:34,349
amino acids and then we have a a

54
00:02:40,680 --> 00:02:37,450
positive control code which is all of

55
00:02:42,839 --> 00:02:40,690
today's amino acids so the protein

56
00:02:45,960 --> 00:02:42,849
variants in each library have a

57
00:02:49,380 --> 00:02:45,970
randomized region of 80 amino acids and

58
00:02:52,979 --> 00:02:49,390
each library has a complexity of greater

59
00:02:55,890 --> 00:02:52,989
than 10 to the 12 unique sequences so

60
00:02:58,979 --> 00:02:55,900
the the power of our approach is that we

61
00:03:03,630 --> 00:02:58,989
planned to compare these four libraries

62
00:03:06,990 --> 00:03:03,640
these four alphabets of this evolving

63
00:03:10,800 --> 00:03:07,000

code and see what propensity they have

64
00:03:12,780 --> 00:03:10,810
for structure and function so the first

65
00:03:15,930 --> 00:03:12,790
thing we did was interrogate for

66
00:03:18,720 --> 00:03:15,940
structure or a proxy proxy for structure

67
00:03:21,449 --> 00:03:18,730
which is the ability to form soluble

68
00:03:23,940 --> 00:03:21,459
protein when expressed in e.coli so I

69
00:03:27,000 --> 00:03:23,950
took about two dozen individual variants

70
00:03:31,830 --> 00:03:27,010
from each library and expressed them and

71
00:03:34,890 --> 00:03:31,840
then used simple soluble versus

72
00:03:39,750 --> 00:03:34,900
insoluble as an idea of whether they can

73
00:03:42,930 --> 00:03:39,760
fold in the cell and a rough conclusion

74
00:03:46,250 --> 00:03:42,940
is that the newer alphabets the 16 and

75
00:03:49,530 --> 00:03:46,260
the 20 were most likely to be expressed

76
00:03:51,629 --> 00:03:49,540
whereas the older alphabets the 5 and

77
00:03:53,879 --> 00:03:51,639
the 9 when they were expressed were more

78
00:03:57,990 --> 00:03:53,889
likely to be soluble which was

79
00:04:00,780 --> 00:03:58,000
interesting so now we had our sometimes

80
00:04:04,410 --> 00:04:00,790
soluble proteins the next thing we

81
00:04:06,809 --> 00:04:04,420
wanted to interrogate was which of these

82
00:04:10,050 --> 00:04:06,819
alphabets or do all of these alphabets

83
00:04:13,259 --> 00:04:10,060
have a propensity for functions that are

84
00:04:15,809 --> 00:04:13,269
essential for life so one of the most

85
00:04:17,190 --> 00:04:15,819
simple functions I can select for

86
00:04:20,490 --> 00:04:17,200
because I kind of wanted to go a little

87
00:04:24,659 --> 00:04:20,500
bit easy on these alphabets is ligand

88
00:04:27,450 --> 00:04:24,669

binding and some essential ligands at

89
00:04:30,870 --> 00:04:27,460
the origin of life as Mark has already

90
00:04:32,480 --> 00:04:30,880
mentioned ATP and gtp so these are

91
00:04:34,730 --> 00:04:32,490
essential in a

92
00:04:37,309 --> 00:04:34,740
currencies for all of life and also

93
00:04:40,279 --> 00:04:37,319
components of RNA which I think we can

94
00:04:44,629 --> 00:04:40,289
agree is somewhat central to origin of

95
00:04:47,839 --> 00:04:44,639
life so the method that I use to

96
00:04:49,850 --> 00:04:47,849
interrogate these libraries for binding

97
00:04:52,189 --> 00:04:49,860
of these cofactors is an in vitro

98
00:04:55,550 --> 00:04:52,199
evolution method again similar to Marc

99
00:04:57,559 --> 00:04:55,560
called mRNA display for those of you who

100
00:05:00,740 --> 00:04:57,569
aren't familiar with it it's a technique

101
00:05:04,850 --> 00:05:00,750
that in which you've physically attached

102
00:05:07,159 --> 00:05:04,860
a protein variant to its own encoding

103
00:05:09,589 --> 00:05:07,169
mRNA and this is really powerful because

104
00:05:12,680 --> 00:05:09,599
it means that you can take a library of

105
00:05:15,200 --> 00:05:12,690
trillions of unique sequences and you

106
00:05:17,210 --> 00:05:15,210
can pluck out a single protein that has

107
00:05:20,899 --> 00:05:17,220
your desired function and because it's

108
00:05:22,760 --> 00:05:20,909
attached to its own code you can decode

109
00:05:27,680 --> 00:05:22,770
it and get its sequence and you can also

110
00:05:28,730 --> 00:05:27,690
propagate it through PCR so aside from

111
00:05:31,180 --> 00:05:28,740
that it's pretty similar to the

112
00:05:35,149 --> 00:05:31,190
technique that Mark spoke about so we

113
00:05:39,290 --> 00:05:35,159

generate our RNA protein fusions we

114
00:05:43,339 --> 00:05:39,300
introduce them to some immobilized ATP

115
00:05:45,950 --> 00:05:43,349
and gtp ligands and then only the

116
00:05:50,120 --> 00:05:45,960
fusions that can be competitively eluted

117
00:05:53,270 --> 00:05:50,130
go on to form the precursor for the next

118
00:05:56,749 --> 00:05:53,280
cycle and you assume that once you have

119
00:06:02,779 --> 00:05:56,759
binding you see you observe increased

120
00:06:05,510 --> 00:06:02,789
enrichment after every round so I first

121
00:06:08,959 --> 00:06:05,520
performed this with our control alphabet

122
00:06:11,450 --> 00:06:08,969
of the extant 20 amino acids and you can

123
00:06:13,999 --> 00:06:11,460
see along the bottom here the the number

124
00:06:17,270 --> 00:06:14,009
of rounds and then the percent of

125
00:06:19,610 --> 00:06:17,280
fusions that are selected per round up

126
00:06:21,680 --> 00:06:19,620
on the y axis and this was really

127
00:06:24,409 --> 00:06:21,690
encouraging to see because after every

128
00:06:27,350 --> 00:06:24,419
round we saw increased enrichment for

129
00:06:29,930 --> 00:06:27,360
ATP and gtp binding you can see that it

130
00:06:33,080 --> 00:06:29,940
is significantly above a negative

131
00:06:35,870 --> 00:06:33,090
control that I did but this was my

132
00:06:37,870 --> 00:06:35,880
positive control and I didn't know how

133
00:06:41,810 --> 00:06:37,880
the other alphabets were going to behave

134
00:06:44,040 --> 00:06:41,820
five amino acids awfully few to expect

135
00:06:45,629 --> 00:06:44,050
much out of so I

136
00:06:48,540 --> 00:06:45,639
I repeated this experiment with

137
00:06:51,779 --> 00:06:48,550
identical conditions for the three other

138
00:06:55,230 --> 00:06:51,789

ancient alphabets and was really

139
00:06:58,770 --> 00:06:55,240
surprised to see that all three of the

140
00:07:02,969 --> 00:06:58,780
reduced amino acid alphabets yielded ATP

141
00:07:06,659 --> 00:07:02,979
and GTP binders after an average of five

142
00:07:08,189 --> 00:07:06,669
rounds so this was it may not have

143
00:07:10,830 --> 00:07:08,199
enriched quite as high for the five

144
00:07:13,680 --> 00:07:10,840
library but that's definite enrichment

145
00:07:16,050 --> 00:07:13,690
and this was a repeatable result so

146
00:07:19,589 --> 00:07:16,060
really excited about this so the next

147
00:07:22,920 --> 00:07:19,599
stage was to find out what these

148
00:07:25,020 --> 00:07:22,930
populations of sequences are so I picked

149
00:07:28,980 --> 00:07:25,030
around from each of the experiments for

150
00:07:30,809 --> 00:07:28,990
deep sequencing to glean a little more

151
00:07:33,209 --> 00:07:30,819
information from this experiment I

152
00:07:36,149 --> 00:07:33,219
repeated all of these rounds and then

153
00:07:38,279 --> 00:07:36,159
after I generated the fusions I split

154
00:07:41,909 --> 00:07:38,289
them into four different conditions so

155
00:07:45,600 --> 00:07:41,919
the first has ATP and gtp pulled

156
00:07:49,499 --> 00:07:45,610
together as my typical selection round

157
00:07:52,260 --> 00:07:49,509
was and then I did ATP by itself gtp by

158
00:07:56,519 --> 00:07:52,270
itself and then ain't no ligand control

159
00:07:59,999 --> 00:07:56,529
as my negative control I had the kind

160
00:08:03,450 --> 00:08:00,009
help of celia blanco and irene chen who

161
00:08:06,570 --> 00:08:03,460
i think here in analyzing this data and

162
00:08:09,450 --> 00:08:06,580
the metric we came up as a proxy for

163
00:08:12,360 --> 00:08:09,460

binding affinity was relative enrichment

164
00:08:15,420 --> 00:08:12,370
so we considered an individual sequence

165
00:08:17,790 --> 00:08:15,430
to be enriched if it had more copies

166
00:08:20,010 --> 00:08:17,800
after selection than it had prior to

167
00:08:22,800 --> 00:08:20,020
selection and we considered it to show

168
00:08:25,019 --> 00:08:22,810
relative enrichment if this rep

169
00:08:27,930 --> 00:08:25,029
enrichment was greater than this

170
00:08:30,420 --> 00:08:27,940
enrichment for the negative control so

171
00:08:32,370 --> 00:08:30,430
we would expect the values for relative

172
00:08:38,459 --> 00:08:32,380
enrichment to be greater than one to

173
00:08:40,709 --> 00:08:38,469
show real specific binding so this is an

174
00:08:42,130 --> 00:08:40,719
awful lot of information but I can walk

175
00:08:44,830 --> 00:08:42,140
you through it

176
00:08:49,000 --> 00:08:44,840
each of these circles represents a

177
00:08:52,090 --> 00:08:49,010
unique sequence on the y-axis we have

178
00:08:54,460 --> 00:08:52,100
relative enrichment on a log scale along

179
00:08:56,470 --> 00:08:54,470
the bottom we have the three different

180
00:09:00,520 --> 00:08:56,480
selection conditions per library

181
00:09:03,640 --> 00:09:00,530
color-coded up here the line here is the

182
00:09:05,110 --> 00:09:03,650
threshold for relative enrichments

183
00:09:07,720 --> 00:09:05,120
everything above this is showing

184
00:09:10,990 --> 00:09:07,730
specific binding everything below we

185
00:09:13,450 --> 00:09:11,000
don't want to think about the median

186
00:09:16,450 --> 00:09:13,460
value is shown in black and straight

187
00:09:19,480 --> 00:09:16,460
away if we look at it we can see this

188
00:09:23,470 --> 00:09:19,490

trend of increased propensity for

189
00:09:26,800 --> 00:09:23,480
binding and increased affinity of

190
00:09:30,160 --> 00:09:26,810
binding with the increased complexity of

191
00:09:33,550 --> 00:09:30,170
the alphabet but there's a plateau after

192
00:09:36,580 --> 00:09:33,560
you reach the 16 amino acid alphabet

193
00:09:40,150 --> 00:09:36,590
which suggests that those final four

194
00:09:41,890 --> 00:09:40,160
amino acids are adding a lot for this

195
00:09:45,810 --> 00:09:41,900
particular function in this particular

196
00:09:49,570 --> 00:09:45,820
selection if we look at the five library

197
00:09:52,450 --> 00:09:49,580
those medians are really low there's

198
00:09:54,550 --> 00:09:52,460
still some binding and the the best

199
00:09:56,440 --> 00:09:54,560
binders yeah they're showing some

200
00:10:00,250 --> 00:09:56,450
relative enrichment but if we look at

201
00:10:02,620 --> 00:10:00,260
the sixteen for ATP it is several orders

202
00:10:06,670 --> 00:10:02,630
of magnitude better than the five and

203
00:10:10,090 --> 00:10:06,680
that was really exciting to see I'm

204
00:10:13,600 --> 00:10:10,100
gonna zoom in now on the the best six

205
00:10:19,060 --> 00:10:13,610
binders from each library from each

206
00:10:21,010 --> 00:10:19,070
alphabet and again this this really

207
00:10:23,650 --> 00:10:21,020
pronounces the the difference and

208
00:10:25,780 --> 00:10:23,660
relative enrichment between the

209
00:10:28,060 --> 00:10:25,790
alphabets so you can see that you have a

210
00:10:30,700 --> 00:10:28,070
relative enrichment between one and ten

211
00:10:35,860 --> 00:10:30,710
for the five and all the way up to over

212
00:10:38,890 --> 00:10:35,870
a thousand for the sixteen but this is

213
00:10:44,530 --> 00:10:38,900

480p and this pattern isn't as evident

214
00:10:46,080 --> 00:10:44,540
for gtp-binding so the the final thing I

215
00:10:48,540 --> 00:10:46,090
wanted to do

216
00:10:50,790 --> 00:10:48,550
I wanted to talk about is to talk about

217
00:10:53,190 --> 00:10:50,800
the specificity of binding as I

218
00:10:56,730 --> 00:10:53,200
mentioned I performed the selection with

219
00:10:59,630 --> 00:10:56,740
a mixture of ligands mostly because I

220
00:11:01,890 --> 00:10:59,640
couldn't be bothered doing it twice and

221
00:11:04,260 --> 00:11:01,900
but it's quite striking if you look

222
00:11:07,410 --> 00:11:04,270
again at just these top six selected

223
00:11:09,570 --> 00:11:07,420
binders alphabet you can see that for

224
00:11:11,910 --> 00:11:09,580
the five and the nine there isn't really

225
00:11:14,100 --> 00:11:11,920
much discrimination between bindings or

226
00:11:17,210 --> 00:11:14,110
the ratio of ATP and gtp is on the y

227
00:11:20,550 --> 00:11:17,220
axis here so there's some really sloppy

228
00:11:22,920 --> 00:11:20,560
nonspecific binding happening whereas if

229
00:11:25,380 --> 00:11:22,930
you look at the sixteen or the twenty

230
00:11:28,230 --> 00:11:25,390
there's much more of a bias one way or

231
00:11:31,190 --> 00:11:28,240
the other towards ATP or GTP for the

232
00:11:35,190 --> 00:11:31,200
twenty and this suggests that as the the

233
00:11:37,380 --> 00:11:35,200
chemistry's of the genetic code get more

234
00:11:39,330 --> 00:11:37,390
complex they are more able to

235
00:11:41,280 --> 00:11:39,340
distinguish between the ligands that

236
00:11:43,260 --> 00:11:41,290
they're binding which is of course a

237
00:11:48,150 --> 00:11:43,270
function that is crucial for modern

238
00:11:51,060 --> 00:11:48,160

biology or any biology so I would like

239
00:11:54,120 --> 00:11:51,070
to go to some really quick take-home

240
00:11:56,850 --> 00:11:54,130
points we managed to find binders from

241
00:11:58,350 --> 00:11:56,860
the five amino acid alphabet and that is

242
00:12:01,680 --> 00:11:58,360
something we never expect to see and

243
00:12:03,660 --> 00:12:01,690
we're really excited about that we

244
00:12:05,850 --> 00:12:03,670
observed that the relative enrichments

245
00:12:08,520 --> 00:12:05,860
the proxy for binding increased with

246
00:12:11,220 --> 00:12:08,530
library complexity the best binders were

247
00:12:13,560 --> 00:12:11,230
from the sixteen library not from the

248
00:12:16,650 --> 00:12:13,570
twenty and the most selective binders

249
00:12:18,990 --> 00:12:16,660
were from the sixteen and the twenty so

250
00:12:21,660 --> 00:12:19,000
I think I can tentatively and make the

251
00:12:25,200 --> 00:12:21,670
cheekily conclude that under some very

252
00:12:27,540 --> 00:12:25,210
specific conditions hypothetical early

253
00:12:30,960 --> 00:12:27,550
genetic codes may actually rival today's

254
00:12:32,310 --> 00:12:30,970
Universal code with that I'd like to

255
00:12:38,870 --> 00:12:32,320
acknowledge people who've worked on this

256
00:12:45,600 --> 00:12:41,880
so we have time for a few questions yes

257
00:12:45,610 --> 00:13:25,320
[Music]

258
00:13:34,590 --> 00:13:27,540
that you've identified a stone in my

259
00:13:38,220 --> 00:13:34,600
shoe yes the the twin the 9 did take

260
00:13:41,280 --> 00:13:38,230
longer and I repeated it and repeated it

261
00:13:44,520 --> 00:13:41,290
and it still would consistently take a

262
00:13:48,780 --> 00:13:44,530
little longer I don't have a good answer

263
00:13:51,660 --> 00:13:48,790

because when we did get binders they are

264
00:13:56,040 --> 00:13:51,670
identify early I identify ibly better

265
00:13:58,710 --> 00:13:56,050
than the 5 one caveat one technical

266
00:14:00,210 --> 00:13:58,720
caveat is that the complexity of the 9

267
00:14:02,730 --> 00:14:00,220
library at the beginning of the

268
00:14:05,670 --> 00:14:02,740
experiment was higher than the other

269
00:14:08,490 --> 00:14:05,680
three libraries so that is something

270
00:14:10,650 --> 00:14:08,500
that we need to correct for when we

271
00:14:13,080 --> 00:14:10,660
publish this and it might be that that

272
00:14:14,970 --> 00:14:13,090
actually explains for that extra lag

273
00:14:17,400 --> 00:14:14,980
before we saw enrichment I just need it

274
00:14:20,250 --> 00:14:17,410
had more sequences to pair through

275
00:14:31,210 --> 00:14:20,260
before it went right up we have time for

276
00:14:39,100 --> 00:14:34,759
these experiments when they finally did

277
00:14:45,079 --> 00:14:39,110
some analysis it turned out that the

278
00:14:47,720 --> 00:14:45,089
interactions um we're in the process of