Схема следующая: для всех 24 файлов(12 из Контроля, 12 из Группы без гиппокампа) ищутся паттерны. Получается 24 набора паттернов. потом для каждого набора файла x, для каждого паттерна из этого набора, смотрим как входит этот паттерн в данные из файла y. Возможно несколько способов подсчета числа «матчей» паттернов из файла x в файле y:

- для каждого паттерна из x, если он(паттерн) длиннее $MinPat$, добавляем к числу матчей *число встреч* этого паттерна в файле y(frequent).
- для каждого паттерна из x, если он(паттерн) длиннее $MinPat$, добавляем к числу матчей *единичку*, если этот паттерна встречается в y(boolean).

Тоесть надо сделать 2 выбора: минимальная длинна паттерна; добавлять число встреч паттерна, или 1, если true, и 0 если false.

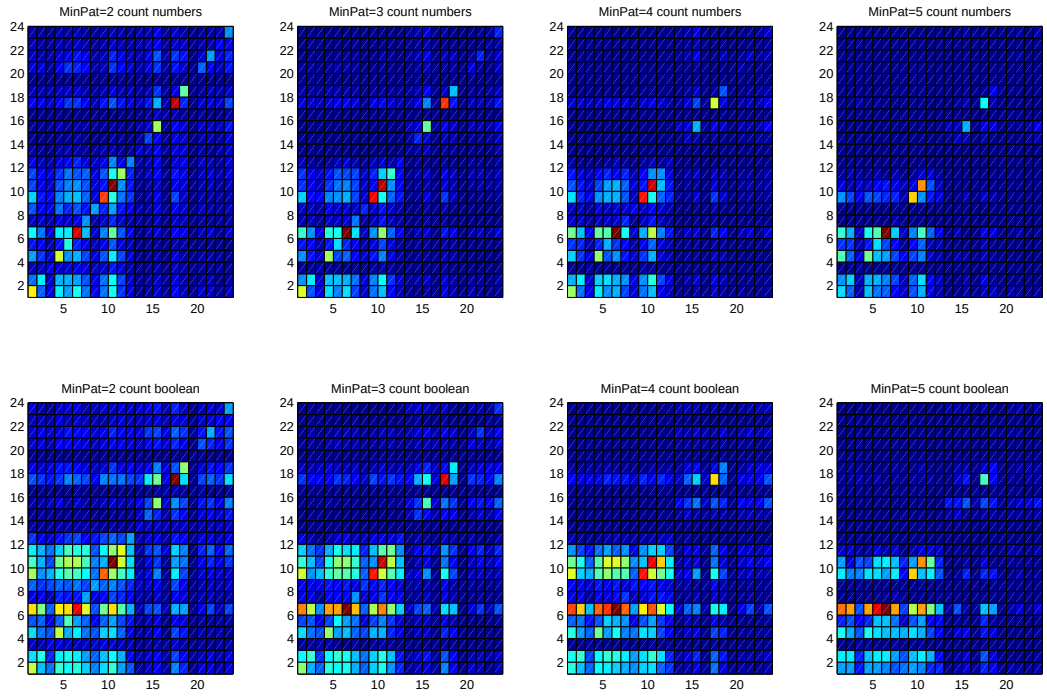

Рис. 1: Матчинги файлов для разных способов подсчета матчингов. Для каждого квадрата:первые 12 ячеек — контроль, вторые 12 — без гиппокампа. По горизонтали откладывается из какого файла берутся паттерны, по вертикали откладывается в каком файле они ищутся. Т.о. правый нижний квадрант показывает «какие паттерны нормальных мышей мы нашли у безгиппокампных», левый верхний квадрант говорит «какие паттерны безгиппокампных мышей мы наши у нормальных». Заметьте, что вообще говоря, здесь не должно быть симметрии, и это иногда можно наблюдать.

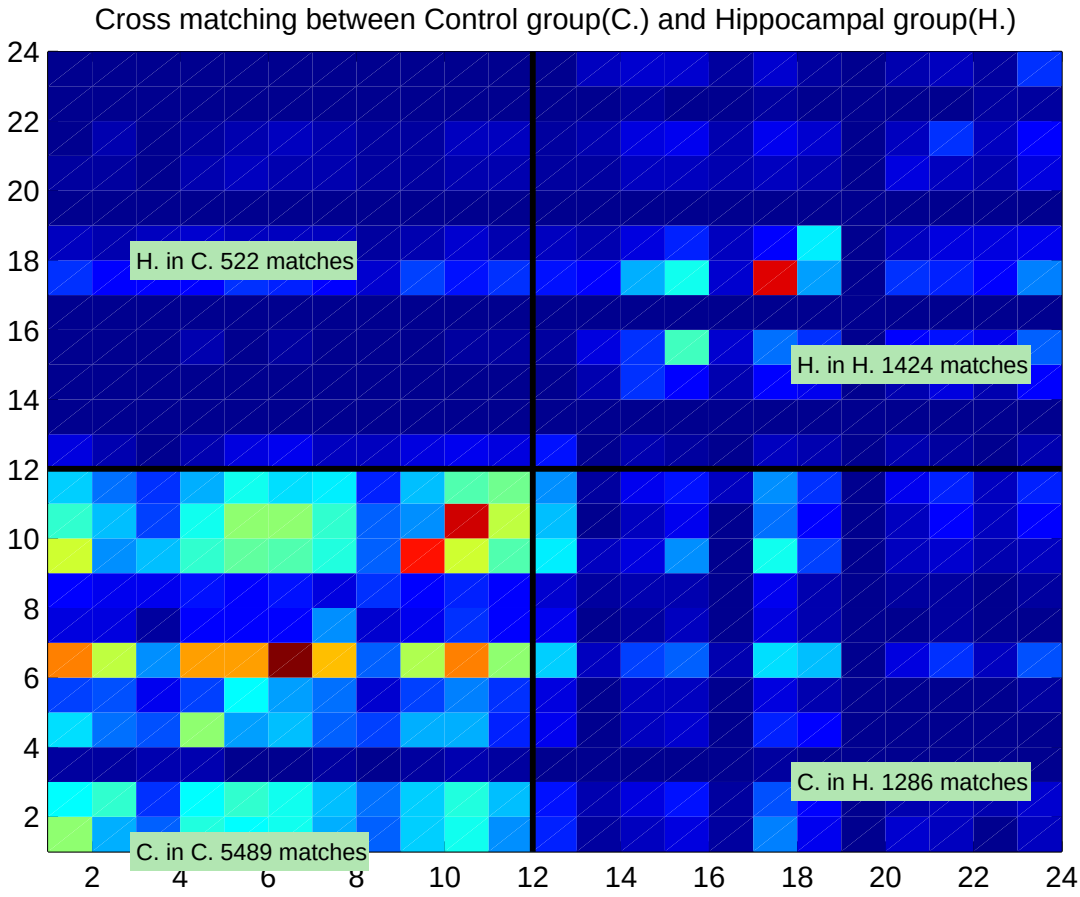

Рис. 2: Более крупно для $MinPat = 2$, boolean

Чисто визуально нравится больше всего случай $MinPat = 2$+boolean(на практике он самый лучший, см. таблицу 1), так как в нем хорошо видны внутренние корреляции групп. Таким образом, действительно «булевский» матчинг более робастный, как и было написано в статье Ступа. Отмечу, что двухвыборочный критерий Уилкоксона(ranksum) и двухвыборочный критерий Колмогорова Смирнова хорошо срабатывают: группы левого нижнего и правого верхнего признаны взятыми из разных распределений. Подгруппы внутри этих квадрантов признаются взятыми из одного распределения при $p = 0.01$.

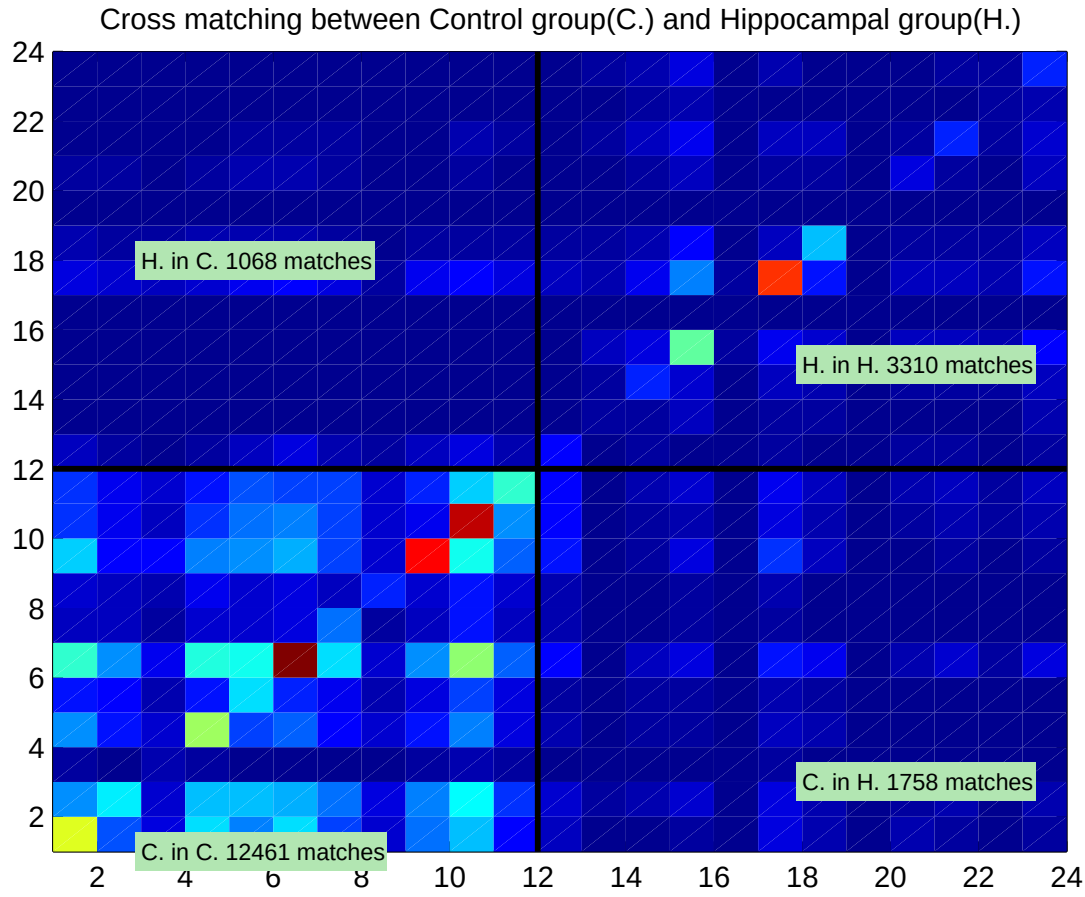

Рис. 3: Более крупно для $MinPat = 2$, frequent

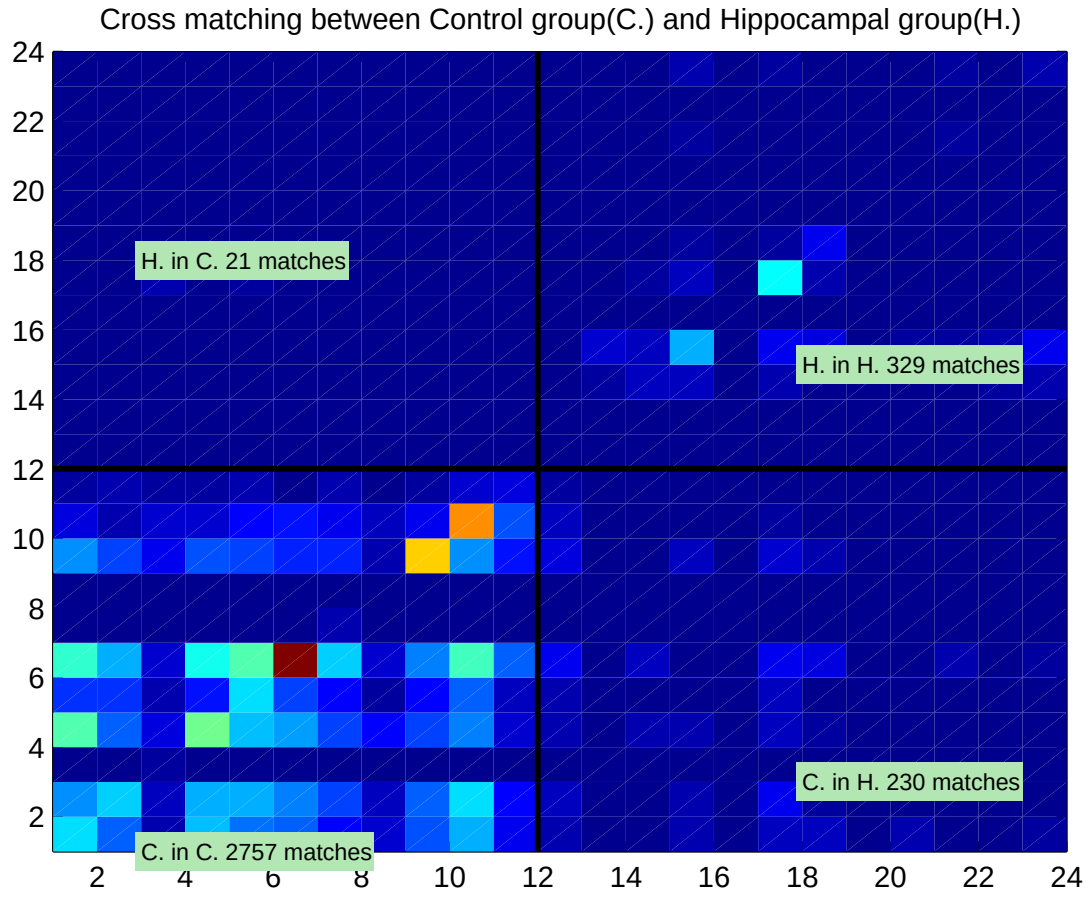

Рис. 4: Более крупно для $MinPat = 4$, frequent

	1	2	3	4	1b	2b	3b	4b
k=1	58%	72%	71%	55%	87%	87%	84%	61%
k=3	53%	59%	58%	56%	69%	73%	59%	49%

Таблица 1: Процент правильных ответов классификации. Первые 4 колонки соответствуют частотному матчингу паттернов соответствующей минимальной длины(подсчет вхождений). Вторые 4 колонки соответствуют булевскому матчингу паттернов соответствующей минимальной длины(подсчет фактов вхождения). P-Паттерны

	1	2	3	1b	2b	3b
k=1	46%	55%	50%	62%	51%	50
k=3	49%	69%	53%	67%	55%	50%

Таблица 2: Процент правильных ответов классификации. Первые 3 колонки соответствуют частотному матчингу паттернов соответствующей минимальной длины(подсчет вхождений). Вторые 3 колонки соответствуют булевскому матчингу паттернов соответствующей минимальной длины(подсчет фактов вхождения). T-Паттерны.

Теперь возьмем метод kNN и проведем схему leave-one-out. Т.е. берем для «обучения» 5 объектов из каждого класса оставляем для теста по одному объекту из класса. Результаты для разным способов матчинга и значений в таблице.

Причем, мне кажется что результат не такой уж тривиальный. По гистограммам не так уж все ясно(самыне нижние рисунки). А матчинг Т-Пттернов дает результаты на порядок хуже. Заметьте, что только, на рисунке 5 видная кое-какая связь и то это связанно исключительно с тем, что данный контрольной группы длиннее, поэтому бóльшую часть вносят короткие паттерны, которые находятся и в шуме. Ну а так как файлы контрольной группы длиннее, то и матчингов этих больше. При других параметров выборки неотличимы, если сравнивать Т-Паттерны. Качество классификации вообще никакое(см.табл. 2).

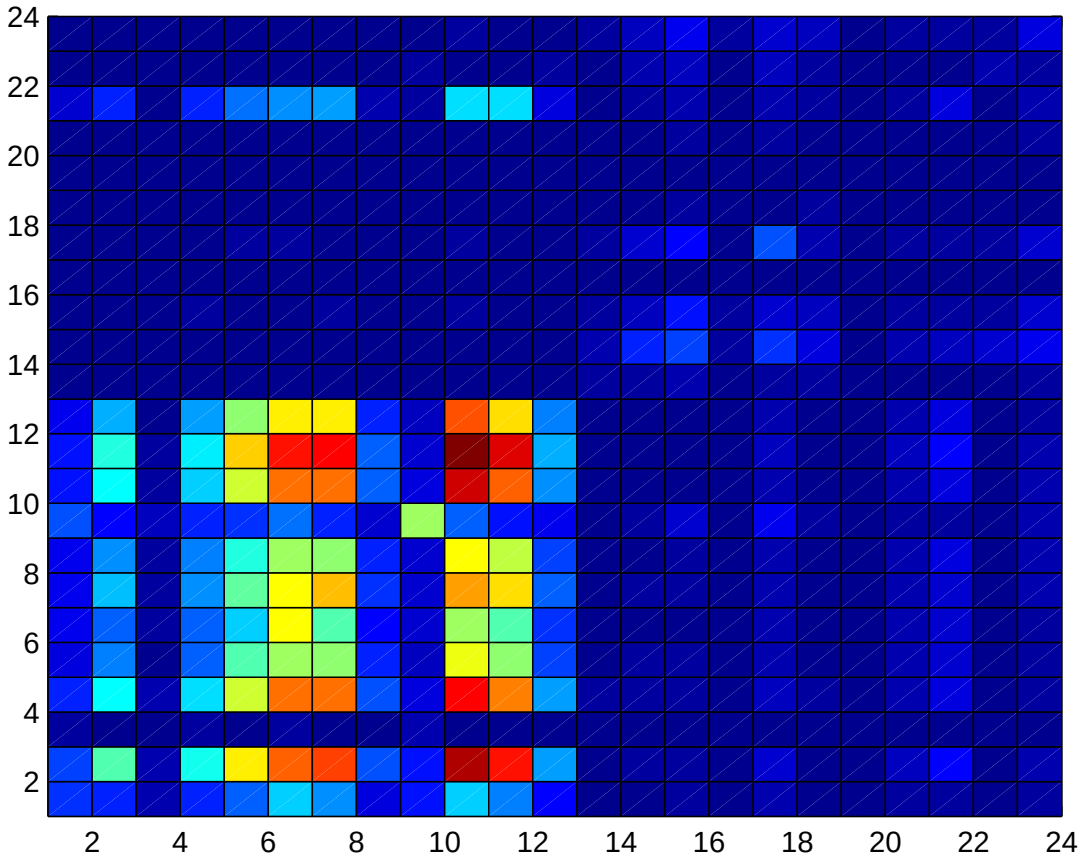

Рис. 5: Более крупно для $MinPat = 2$, frequent. T-Паттерны

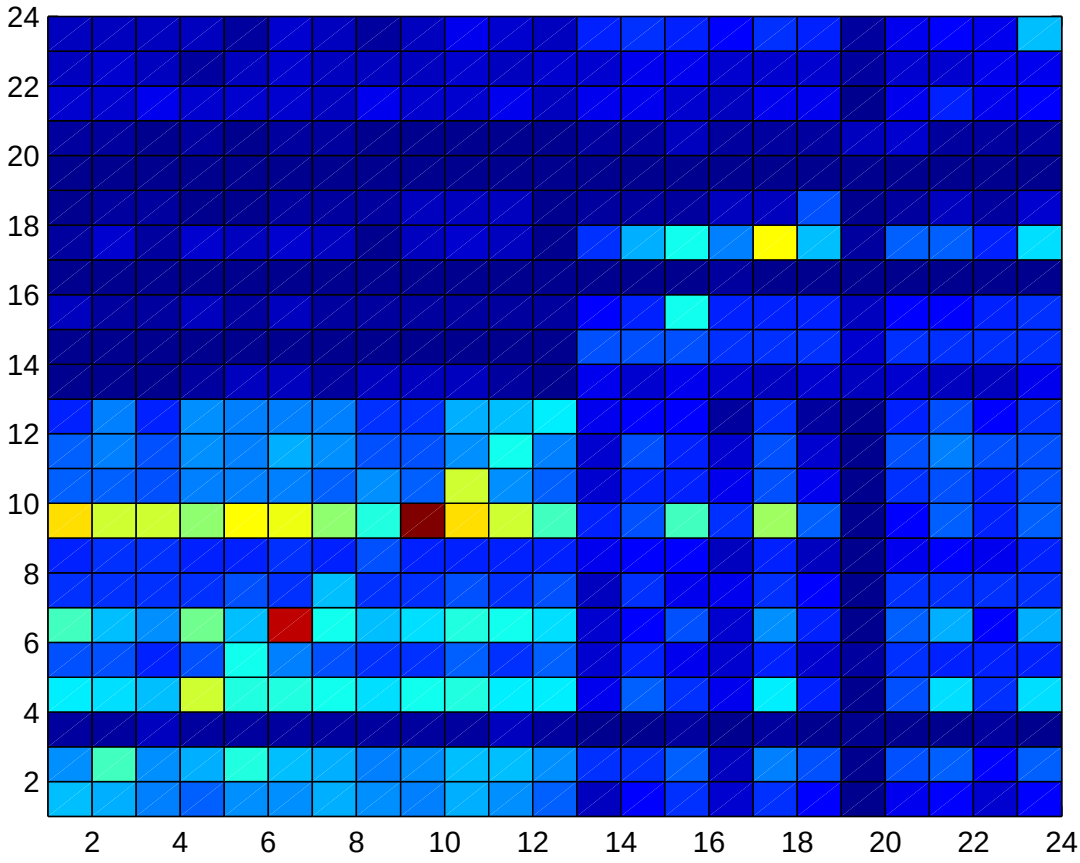

Рис. 6: Более крупно для $MinPat = 2$, boolean. T-Паттерны

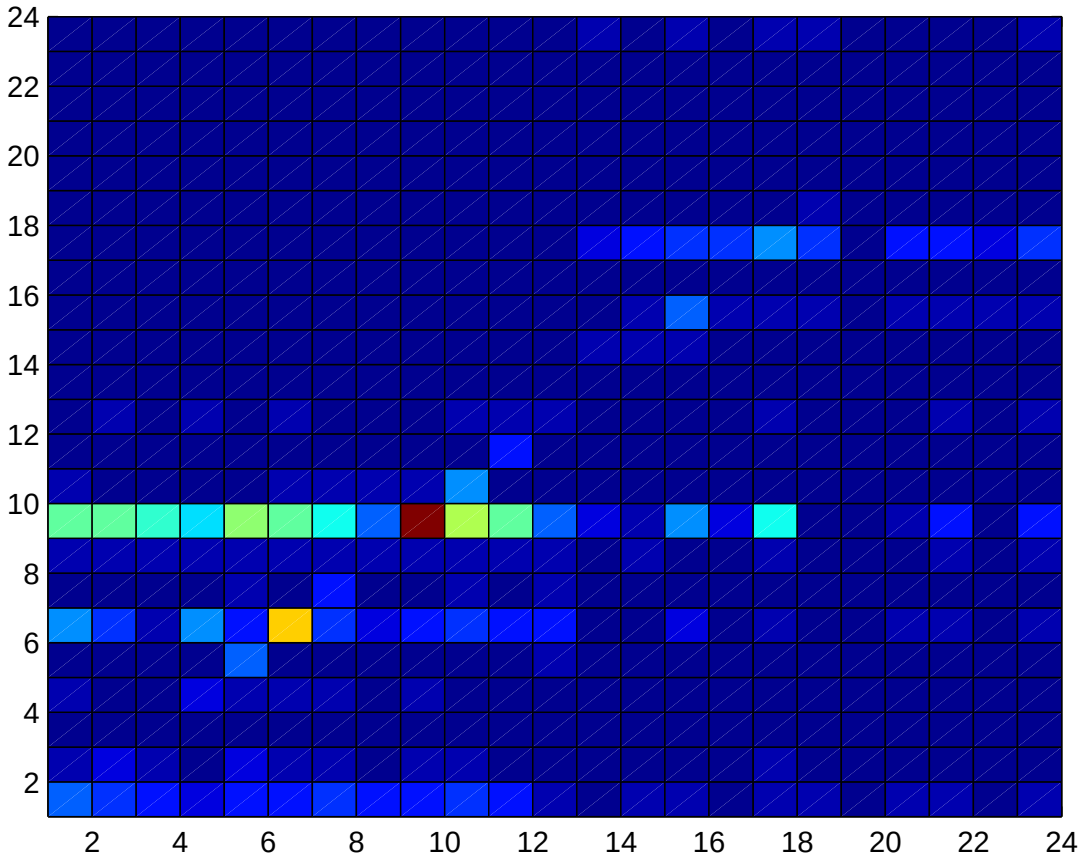

Рис. 7: Более крупно для $MinPat = 4$, boolean. T-Паттерны

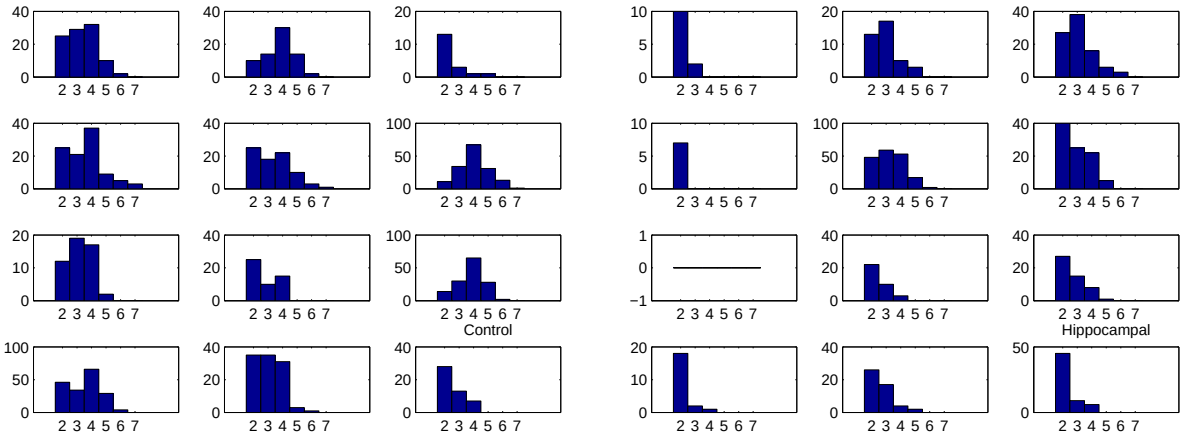

Рис. 8: Гистограмма распределения длин P-Паттернов в контрольной группе(слева) и группе мышей без гиппокампа(справа).

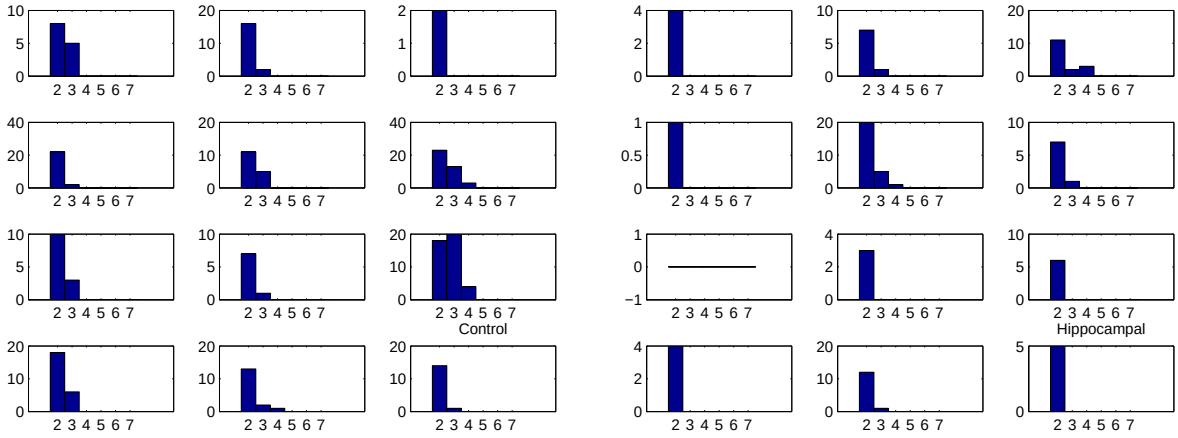

Рис. 9: Гистограмма распределения длин Т-Паттернов в контрольной группе(слева) и группе мышей без гиппокампа(справа).